Statistical Estimation in High Dimension, Sparsity and Oracle Inequalities

Karim Lounici

UNIVERSITE PARIS.DIDEROT (Paris 7)
UFR de Mathématiques

Thèse

pour obtenir le titre de

Docteur de l'Université PARIS.DIDEROT (Paris 7)

Spécialité : Mathématiques Appliquées

présentée par

Karim LOUNICI

Estimation Statistique En Grande Dimension, Parcimonie Et Inégalités D'Oracle

---

Sous la direction d' **Alexandre TSYBAKOV**

Soutenue publiquement le 24 novembre 2009, devant le jury composé de :

M.	**Francis**	**BACH**	INRIA	(Rapporteur)
M.	**Stéphane**	**BOUCHERON**	Université Paris.Diderot (Paris 7)	
M.	**Arnak**	**DALALYAN**	CERTIS - Ecole des Ponts	
M.	**Pascal**	**MASSART**	Université Paris-Sud 11	
M.	**Alexandre**	**TSYBAKOV**	Université Paris 6 - CREST	
M.	**Nicolas**	**VAYATIS**	Ens Cachan	

Remerciements

Mes premiers remerciements vont à Alexandre “Sacha” Tsybakov pour son attention constante lors de ces trois dernières années. Il m'a apporté énormément grâce à son encadrement scientifique, ses qualités humaines ainsi que par toutes les opportunités et les rencontres qu'il a rendues possible. Sacha, j'ai pour toi la plus profonde gratitude. Merci ! Je tiens également à remercier Nicolas Vayatis pour son soutien et ses conseils avisés tout au long de cette thèse.

Francis Bach et Vladimir Koltchinskii ont gracieusement accepté d'examiner cette thèse. C'est pour moi un grand honneur d'être évalué par d'aussi éminents chercheurs.

Je suis très honoré de pouvoir compter parmi les membres de mon jury Pascal Massart dont j'admire les travaux. Au cours de cette thèse, Stéphane Boucheron et Arnak Dalalyan ont toujours pris le temps de répondre à mes questions et de me prodiguer d'excellents conseils. Je les remercie pour leur gentillesse et me réjouis de leur participation à mon jury.

Je tiens à exprimer ma gratitude envers Dominique Picard, Gérard Kerkyacharian et Lucien Birgé pour la qualité de leurs enseignements qui ont accru mon engouement pour les statistiques mathématiques.

Je remercie Christian Robert et le laboratoire de Statistiques du CREST pour m'avoir accueilli pendant deux années dans d'excellentes conditions.

Je souhaite également remercier Pierre Alquier, Patrice Bertail, Gérard Biau, Christina Butucea, Laurent Cavalier, Christophe Chesneau, Stéphane Gaïffas, Emmanuelle Gautherat, Eric Gautier, Mohamed Hebiri, Guillaume Lecué, Erwan Le Pennec, Katia Meziani, Thanh Mai Pham Ngoc, Massimiliano Pontil, Sara van de Geer, Christophe Pouet, Vincent Rivoirard, Angelika Rohde, Etienne Roquain, Mathieu Rosenbaum, Joseph Salmon et Gilles Stoltz pour les discussions plus ou moins formelles que nous avons eues lors d'un séminaire, colloque ou d'une simple pause-café.

Je salue les thésards des bureaux 5C9 et F14 pour les bons moments passés ensemble : Julien, Christophe, Luca, François, Maxime, Pierre, Hubert, Gregorio, Marc, Eun Jung, Aude, Lionel et tous autres.

Un grand merci à l'équipe administrative du LPMA et de Paris 7 pour son efficacité et tout particulièrement à Valérie Juvé et Michelle Wasse qui m'ont toujours aidé à résoudre toutes les tracasseries administratives.

Je remercie également Lionel Alberti et Pierre Clare pour leur amitié.

Enfin, je réserve ces dernière lignes à ma famille : mes parents, mon frère Amar et ma soeur Nawel pour leur soutien et leurs encouragements.

TABLE DES MATIÈRES

Chapitre 1

Introduction

Dans cette thèse nous traitons deux sujets. Le premier sujet concerne l'apprentissage statistique en grande dimension, i.e., les problèmes où le nombre de paramètres potentiels est beaucoup plus grand que le nombre de données à disposition. Dans ce contexte, l'hypothèse généralement adoptée est que le nombre de paramètres intervenant effectivement dans le modèle est petit par rapport au nombre total de paramètres potentiels et aussi par rapport au nombre de données. Cette hypothèse est appelée "*sparsity assumption*". Nous étudions les propriétés statistiques de deux types de procédures :

- les procédures basées sur la minimisation du risque empirique muni d'une pénalité l_1 sur l'ensemble des paramètres potentiels.
- les procédures à poids exponentiels.

Le second sujet que nous abordons concerne l'étude de procédures d'agrégation dans un modèle de densité. Notre but est d'établir des inégalités d'oracle pour la norme L^π, $1 \leqslant \pi \leqslant \infty$. Nous proposons ensuite une application à l'estimation minimax et adaptative en la régularité de densité. La construction d'une procédure d'estimation minimax nécessite la connaissance de la régularité de la densité. Or cette quantité est très souvent inconnue en pratique. L'approche que nous adoptons dans ce cas est de calculer tous les estimateurs pour une grille suffisamment fine de valeurs du paramètre inconnu. Puis nous appliquons nos procédures d'agrégation pour construire un nouvel estimateur qui se comportera aussi bien que le meilleur estimateur de base. Nous construisons ainsi un estimateur minimax et adaptatif en la régularité de la densité à estimer.

1.1 Grande dimension, sparsité

1.1.1 Problème d'optimisation stochastique général

De nombreux problèmes en statistique peuvent se reformuler sous la forme d'un problème d'optimisation stochastique. Soit $(Z, \mathcal{A})$ un espace mesurable, $\Theta \subset \mathbb{R}^M$ un sous-ensemble de $\mathbb{R}^M$ avec $M \geqslant 2$. Soit Z une variable aléatoire à valeurs dans $\mathcal{Z}$. Soit $Q : \mathcal{Z} \times \Theta \to \mathbb{R}^+$ une fonction de perte telle que $Q(z, \cdot)$ est convexe pour tout $z \in \mathcal{Z}$. Le risque intégré est défini par

$$R(\theta) = \mathbb{E}(Q(Z, \theta)).$$

Soit

$$\Theta^* = \arg\min_{\theta \in \mathbb{R}^M} R(\theta)$$

l'ensemble des minimiseurs de R. Si l'ensemble Θ^* n'est pas réduit à un élément se pose la question délicate de l'identifiabilité du minimiseur θ^* d'intérêt. La réponse dépend à la fois de conditions sur la loi de la variable aléatoire Z, de la fonction de perte Q, de la procédure d'estimation employée (poids exponentiels, mininisation du risque empirique avec pénalité l_1) et enfin de la parcimonie (nombre de composantes non nulles) du vecteur θ^* recherché.

Nous verrons ainsi que sous certaines hypothèses sur la loi de Z, les vecteurs $\theta^* \in \Theta^*$ parcimonieux, i.e., avec peu de composantes non nulles, peuvent être reconstruits par les estimateurs Lasso et Dantzig Selector. Pour l'instant, nous faisons l'hypothèse simplificatrice que le minimiseur de R le plus parcimonieux, i.e., avec le nombre minimal de composantes non nulles, est **unique**. Soit donc $\theta^* \in \Theta^*$ le vecteur avec le nombre minimal de composantes non nulles.

Dans un problème statistique donné, nous pouvons considérer que la quantité d'intérêt est $T(\theta^*)$ où $T : \Theta \to \mathcal{T}$ est une application mesurable et $\mathcal{T}$ est un sous-ensemble de $\mathbb{R}^m$ où $m \geqslant 1$. Comme la loi de Z est inconnue, la quantité $T(\theta^*)$ n'est pas accessible directement. Néanmoins nous disposons d'un échantillon i.i.d. $\mathbb{Z}_n = (Z_1, \ldots, Z_n)$ de variables Z_i à valeurs dans $\mathcal{Z}$ et de même loi que Z. Nous allons donc construire un estimateur $\hat{\theta}$ à partir de cet échantillon tel que $T(\hat{\theta})$ estime $T(\theta^*)$.

Nous nous intéresserons dans la suite à trois types de résultats distincts.

- **Problème de prédiction**. Dans ce cas $\mathcal{T} = \mathbb{R}$ et $T(\theta) = R(\theta)$. Le but est de construire un estimateur $\hat{\theta}$ à partir de $\mathbb{Z}_n$ tel que

$$\mathbb{P}\left(R(\hat{\theta}) \leqslant R(\theta^*) + \Delta(s, n, M, \epsilon)\right) \geqslant 1 - \epsilon, \quad 0 < \epsilon < 1 \tag{1.1}$$

où le terme $\Delta(s, n, M, \epsilon) \geqslant 0$ est raisonnablement petit et $s \in \mathbb{N}$ désigne le nombre de composantes non nulles de θ^*.

– **Problème d'estimation**. Dans ce cas $\mathcal{T} = \Theta$ et $T(\theta) = \theta$. Le but est de construire un estimateur $\hat{\theta}$ à partir de $\mathbb{Z}_n$ tel que

$$\mathbb{P}\left(|\hat{\theta} - \theta^*|_p \leqslant \Delta'(s, n, M, \epsilon, p)\right) \geqslant 1 - \epsilon, \quad \forall 0 < \epsilon < 1 \tag{1.2}$$

où $p \in [1, \infty]$, pour tout vecteur $u = (u_1, \ldots, u_M) \in \mathbb{R}^M$, la norme l_p de u est $|u|_p = \left(\sum_{j=1}^M |u_j|^p\right)^{1/p}$ si $p < \infty$ et $|u|_\infty = \max_{1 \leqslant j \leqslant M} |u_j|$ et le terme $\Delta'(s, n, M, \epsilon, p) \geqslant 0$ est raisonnablement petit.

– **Sélection de variables**. Dans ce cas $\mathcal{T} = \{-1, 0, 1\}^M$ et $T(\theta) = \overrightarrow{\text{sign}}(\theta)$, où pour tout vecteur $\theta \in \mathbb{R}^M$, $\overrightarrow{\text{sign}}(\theta) = (\text{sign}(\theta_1), \ldots, \text{sign}(\theta_M))^T$ et pour tout $t \in \mathbb{R}$,

$$\text{sign}(t) = \begin{cases} 1 & \text{if } t > 0, \\ 0 & \text{if } t = 0, \\ -1 & \text{if } t < 0. \end{cases}$$

Le but est de construire un estimateur $\hat{\theta}$ à partir de $\mathbb{Z}_n$ tel que

$$\mathbb{P}\left(\overrightarrow{\text{sign}}(\hat{\theta}) = \overrightarrow{\text{sign}}(\theta^*)\right) \geqslant 1 - \epsilon, \tag{1.3}$$

où $\epsilon > 0$ est aussi petit que possible.

Ce formalisme est très souple. De nombreux problèmes rentrent dans le cadre formulé précédemment. Nous en présentons quelques exemples ci-dessous. Le cadre stochastique que nous proposons ci-dessus peut être facilement modifié pour traiter le cas où les observations sont indépendantes mais non identiquement distribuées. Nous établissons d'ailleurs dans cette thèse des résultats sous cette hypothèse plus faible d'observations indépendantes.

Nous insistons aussi sur le fait qu'en grande dimension, il est crucial d'obtenir des inégalités oracle qui exploitent la parcimonie du modèle, i.e., telles que les termes $\Delta(s, n, M, \epsilon)$ et $\Delta'(s, n, M, \epsilon, p)$ soient linéaires en s et logarithmiques en M. Nous préciserons ce point dans la section 1.1.3.

1.1.2 Exemples

– **Régression.** Soit $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ où $\mathcal{X}$ est un sous-ensemble de $\mathbb{R}^d$ où $d \geqslant 1$. La variable $Z = (X, Y)$ vérifie

$$Y = f(X) + W,$$

où le bruit W est centré de variance σ^2 et $f : \mathcal{X} \to \mathbb{R}$ est la fonction de régression inconnue. Soit $\mathcal{D} = \{f_1, \ldots, f_M\}$ un dictionnaire de fonctions $f_j : \mathcal{X} \to \mathbb{R}$. Soit Θ un sous-ensemble de $\mathbb{R}^M$. Pour tout $\theta = (\theta_1, \ldots, \theta_M) \in \Theta$, notons $\mathsf{f}_\theta = \sum_{j=1}^M \theta_j f_j$ la combinaison linéaire des élements du dictionnaire $\mathcal{D}$. Soit $l : \mathbb{R} \to \mathbb{R}^+$ une fonction convexe. Nous définissons la fonction de perte $Q : \mathcal{Z} \times \Theta \to \mathbb{R}^+$ par

$$Q(z, \theta) = l(y - \mathsf{f}_\theta(x)),\ z = (x, y)$$

Nous pouvons considérer par exemple la perte quadratique avec $l(x) = x^2$ ou bien la perte de Huber avec $l(x) = \frac{1}{2}x^2 \mathbb{I}_{|x| \leqslant 1} + (|x| - \frac{1}{2})\mathbb{I}_{|x|>1}$.

– **Régression transformée.** Soit $\mathcal{Z} = \mathcal{X} \times \{-1; 1\}$ où $\mathcal{X}$ est un sous-ensemble de $\mathbb{R}^d$. Soit la probabilité conditionnelle

$$\mathbb{P}(Y = 1 | X = x) = \pi(x),$$

où π est une fonction inconnue de la forme

$$\pi(x) = \phi'(f(x)),$$

où la fonction $f : \mathcal{X} \to \mathbb{R}$ est inconnue et la fonction $\phi : \mathbb{R} \to \mathbb{R}$ est connue et différentiable sur $\mathbb{R}$. Nous considérons la fonction de perte

$$Q(z, \theta) = -y\mathsf{f}_\theta(x) + \phi(\mathsf{f}_\theta(x)),\ z = (x, y).$$

Un exemple de fonction ϕ est la fonction logit définie par $\phi(x) = \log(1 + e^x)$. Dans ce cas, nous obtenons le modèle de régression logistique.

– **Estimation de densité.** Dans ce cas $\mathcal{Z} = \mathcal{X}$. Nous supposons que la variable aléatoire $X \in \mathcal{X}$ admet la densité de probabilité inconnue f par rapport à la mesure de Lebesgue. La norme $L^2(\mathcal{X})$ est définie par $\|g\| = (\int_{\mathcal{X}} g(x)^2 dx)^{1/2}$ où $g \in L^2(\mathcal{X})$. Etant donné le dictionnaire $\mathcal{D}$, nous souhaitons trouver la combinaison linéaire f_θ qui minimise

$$\|f - \mathsf{f}_\theta\|^2 = \int_{\mathcal{X}} (f - \mathsf{f}_\theta)^2(x) dx = \|f\|^2 + \|\mathsf{f}_\theta\|^2 - 2\mathbb{E}(\mathsf{f}_\theta(X)).$$

Par conséquent, minimiser $\|f - \mathsf{f}_\theta\|^2$ revient à minimiser la quantité

$$R(\theta) = -2\mathbb{E}(\mathsf{f}_\theta(X)) + \|\mathsf{f}_\theta\|^2.$$

La fonction de perte $Q : \mathcal{X} \times \Theta \to \mathbb{R}^+$ est définie par

$$Q(x, \theta) = -2\mathsf{f}_\theta(x) + \|\mathsf{f}_\theta\|^2.$$

– **Régression linéaire multi-tâches**. Soit $\mathcal{Z} = (\mathcal{X} \times \mathbb{R})^T$ où $\mathcal{X}$ est un sous-ensemble de $\mathbb{R}^M$ et $T \geqslant 2$. La variable $Z = ((X_1, Y_1), \ldots, (X_T, Y_T))$ vérifie

$$Y_t = X_t^T \theta_t^* + W_t, \, \forall 1 \leqslant t \leqslant T,$$

où les variables de bruit W_t sont centrées de variance σ^2 et les vecteurs de paramètres inconnus $\theta_t^* \in \mathbb{R}^M$, $1 \leqslant t \leqslant T$, ont des caractéristiques communes, notamment il est raisonnable dans certaines applications de supposer que l'ensemble des composantes non nulles est le même pour tous les vecteurs θ_t^*, $1 \leqslant t \leqslant T$. Nous donnons dans le chapitre 4 des exemples d'applications où cette hypothèse est naturelle.
Soit $l : \mathbb{R} \to \mathbb{R}^+$ une fonction convexe. Soit $\Theta = \otimes_{t=1}^T \Theta_t$ où $\Theta_t \subset \mathbb{R}^M$, $\forall 1 \leqslant t \leqslant T$. Nous définissons la fonction de perte $Q : \mathcal{Z} \times \Theta \to \mathbb{R}^+$ par

$$Q(z, \theta) = \frac{1}{T} \sum_{t=1}^{T} l(y_t - X_t^T \theta_t), \; z = ((x_t, y_t))_{1 \leqslant t \leqslant T}.$$

1.1.3 Compromis entre performances statistiques et algorithmiques

Les problèmes actuels en biologie ou en informatique nécessitent de traiter des données de très grande dimension. C'est-à-dire que le nombre de paramètres inconnus M est beaucoup plus grand que la taille n de l'échantillon à disposition. Les puces à ADN, par exemple, sont des tableaux contenant les expressions de milliers de gènes mesurées sur quelques dizaines d'individus. Dans ce cas, les ordres de grandeurs sont $M \simeq 10^4$ paramètres inconnus contre une taille d'échantillon $n \lesssim 10^2$.

Pour bien comprendre les difficultés soulevées par ce type de données en grande dimension, considérons le modèle de régression linéaire

$$Y = X\theta^* + W, \tag{1.4}$$

où X est une matrice déterministe de taille $n \times M$, le bruit $W = (W_1, \ldots, W_n)$ est un vecteur de variables gaussiennes i.i.d. $\mathcal{N}(0, \sigma^2)$ et $\theta^* = (\theta_1^*, \ldots, \theta_M^*) \in \mathbb{R}^M$ est le vecteur de paramètres inconnus à estimer à partir des observations (X, Y). L'estimateur des moindres carrés est

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^M} |Y - X\theta|_2^2,$$

où pour tout $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, $|x|_2 = (\sum_{i=1}^n x_i^2)^{1/2}$. Supposons, dans un premier temps, que $M \leqslant n$ et que X est de rang plein. Dans ce cas, l'estimateur des moindres carrés est unique et vaut

$$\hat{\theta} = (X^T X)^{-1} X^T Y = \theta^* + (X^T X)^{-1} X^T W.$$

L'erreur moyenne intégrée de prédiction pour la perte l_2 vaut

$$\frac{1}{n}\mathbb{E}|X(\hat{\theta}-\theta^*)|_2^2 = \sigma^2\frac{M}{n}. \tag{1.5}$$

Supposons maintenant que le vecteur inconnu θ^* est s-sparse, c'est-à-dire que le nombre de composantes non nulles de θ^* est égal à s où $1 \leqslant s \leqslant M$. Si nous connaissons le support $J^* = \{j \,:\, \theta_j^* \neq 0\}$, nous pouvons considérer le modèle réduit suivant :

$$Y = X_{J^*}\overline{\theta}^*_{J^*} + W,$$

où X_{J^*} est une sous-matrice de X de taille $n \times s$ et $\overline{\theta}^*_{J^*} \in \mathbb{R}^s$ est obtenu en ne gardant que les composantes non nulles de θ^*. L'erreur moyenne intégrée de prédiction de l'estimateur des moindres carrés $\hat{\theta}_{J^*}$ pour le modèle réduit vaut dans ce cas

$$\frac{1}{n}\mathbb{E}|X_{J^*}(\hat{\theta}_{J^*} - \overline{\theta^*}_{J^*})|_2^2 = \sigma^2\frac{s}{n}. \tag{1.6}$$

Désormais, nous appellerons $\hat{\theta}_{J^*}$ "estimateur oracle" et le risque (1.6) associé "risque oracle".

Si le vecteur θ^* est s-sparse, alors la connaissance de son support permet de passer d'une erreur de prédiction de $\sigma^2 M/n$ pour l'estimateur des moindres carrés $\hat{\theta}$ à $\sigma^2 s/n$ pour l'estimateur oracle $\hat{\theta}_{J^*}$. Nous constatons sur cet exemple l'importance de l'hypothèse de parcimonie. En effet, si $M = n$ et $s = o(n)$ quand $n \to \infty$, alors l'erreur de prédiction (1.5) de $\hat{\theta}$ est constante et égale à σ^2 pour tout n tandis que l'erreur de prédiction (1.6) de $\hat{\theta}_{J^*}$ tend vers 0 lorsque $n \to \infty$.

En pratique, le support J^* est inconnu. La question naturelle est alors la suivante :

Pouvons-nous construire un estimateur $\hat{\theta}$ qui soit approximativement aussi performant que l'estimateur oracle $\hat{\theta}_{J^*}$ au sens de (1.6) lorsque le support J^* est inconnu et le nombre de paramètres M peut être plus grand que la taille n de l'échantillon à disposition ?

Etonnament, la réponse est oui. Nous pouvons proposer des estimateurs $\hat{\theta}$ tels que

$$\frac{1}{n}\mathbb{E}|X(\hat{\theta}-\theta^*)|_2^2 \leqslant C\sigma^2\frac{s}{n}\log M, \tag{1.7}$$

où $C > 0$. Le seul prix à payer pour la non connaissance du support J^* est la présence du terme logarithmique $\log M$.

Schwarz [93] a proposé l'estimateur BIC dans un contexte de sélection de modèle sur une classe finie de modèles $(m_\gamma)_{\gamma\in\Gamma}$. Cet estimateur est défini comme le minimum du risque

empirique muni de la pénalité $\dim(m_\gamma)\sigma^2(\log n)/n$ où $\dim(m_\gamma)$ désigne la dimension du modèle m_γ. Foster et George [40] propose une formulation plus générale de l'estimateur BIC qui englobe [93] et la variante ci-dessous :

$$\hat{\theta}^{BIC} = \arg\min_{\theta\in\mathbb{R}^M} \left(\frac{1}{n}|Y - X\theta|_2^2 + AM(\theta)\sigma^2 \frac{\log M}{n} \right), \tag{1.8}$$

où $A > 0$ et $M(\theta)$ désigne le nombre de composantes non nulles de θ. Bunea, Tsybakov et Wegkamp [15] ont montré que cet estimateur possède de très bonnes propriétés statistiques pour le problème de prédiction puisqu'il vérifie (1.7). En revanche, le problème de minimisation (1.8) est difficile à résoudre en pratique à cause de la présence du terme de pénalisation $l_0 : M(\theta)$. Il s'agit en effet d'un problème d'optimisation non convexe et donc calculable que pour de très petites valeurs de la dimension, soit M de l'ordre de quelques dizaines.

Cet exemple permet d'appréhender la problématique de la grande dimension, à savoir l'intérêt de construire des estimateurs qui réalisent un compromis acceptable entre performances statistiques au sens de (1.7) et bonnes performances algorithmiques, c'est-à-dire s'assurer que les estimateurs proposés sont calculables pour des valeurs de M grandes.

1.1.4 Estimateurs étudiés

Etant donné l'échantillon $\mathbb{Z}_n$, nous définissons le risque empirique par

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} Q(Z_i, \theta).$$

Dans cette thèse, nous étudions les estimateurs suivants :

- **Lasso :**

$$\hat{\theta}^L = \arg\min_{\theta\in\Theta} \left\{ R_n(\theta) + 2r|\theta|_1 \right\},$$

  où $r = A\sigma\sqrt{\frac{\log M}{n}}$, $A > 0$ et pour tout $\theta = (\theta_1, \ldots, \theta_M) \in \Theta$, $|\theta|_1 = \sum_{j=1}^M |\theta_j|$.

- **Dantzig Selector :**

$$\hat{\theta}^D = \arg\min_{\theta\in\Theta} \left\{ |\theta|_1 \,:\, |\nabla R_n(\theta)|_\infty \leqslant r \right\},$$

  où r est défini ci-dessus et pour tout $\theta = (\theta_1, \ldots, \theta_M) \in \Theta$, $|\theta|_\infty = \max_{1\leqslant j\leqslant M} |\theta_j|$.

- **Group Lasso :** dans ce cas que $\Theta = \otimes_{t=1}^T \Theta_t$ et

$$\hat{\theta}^{GL} = \arg\min_{\theta\in\Theta} \left\{ R_n(\theta) + 2r\sum_{j=1}^{M} \|\theta^j\| \right\},$$

où $r = \frac{2\sigma}{\sqrt{nT}}\sqrt{1+\frac{A\log M}{\sqrt{T}}}$, $A>0$, $\theta^j = (\theta_t^j)_{1\leqslant t\leqslant T} \in \mathbb{R}^T$ et $\|\theta^j\| = (\sum_{t=1}^T (\theta_t^j)^2)^{1/2}$, pour tout $1 \leqslant j \leqslant M$.

– **Mirror averaging :** Fixons $\beta > 0$. Soit P_0 et Π des lois de probabilité sur Θ. Pour toute fonction ψ continue bornée sur Θ la mesure de Gibbs $G_{\beta,\Pi}(\psi)$ sur Θ est la mesure de probabilité admettant la densité $\frac{e^{-\psi/\beta}}{\int_\Theta e^{-\psi/\beta} d\Pi}$ par rapport à la mesure de probabilité Π. Soit la mesure de probabilité

$$\hat{P}_n = \frac{1}{n}\left(P_0 + \sum_{k=1}^{n-1} G_{\beta,\Pi}\left(\sum_{i=1}^{k} Q(Z_i,\cdot)\right)\right)$$

L'estimateur à poids exponentiels avec moyennage est défini par

$$\hat{\theta}^{MA} = \int_{\theta\in\Theta} \theta d\hat{P}_n(\theta).$$

L'estimateur $\hat{\theta}^{MA}$ peut s'interpréter d'un point de vue bayésien comme étant la moyenne par rapport à la loi a posteriori $\hat{P}_n$ correspondant à la loi a priori Π. Les estimateurs Lasso, Dantzig Selector et Group Lasso sont les solutions de problèmes de minimisation convexe. En grande dimension, ces problèmes peuvent admettre plusieurs solutions distinctes, ce qui complique l'analyse de ces estimateurs.

1.1.5 Etat de l'art

Le Lasso.

L'estimateur Lasso apparaît peut-être pour la première fois dans la littérature statistique dans Franck et Friedman [41] comme un cas particulier de l'estimateur Bridge. Cet estimateur a été introduit ensuite en théorie du signal par [21] sous le nom de *Basis Pursuit De-noising*. Le nom "Lasso" a été introduit par Tibshirani [94]. Il existe une importante littérature sur cet estimateur. Les premiers résultats obtenus concernent le cas où la dimension M est fixée et $n \to \infty$. Knight et Fu [61] établissent la consistance de l'estimateur Lasso pour le problème d'estimation et pour des asymptotiques spécifiques du paramètre $\lambda = \lambda_n$. Zou [117] démontre la consistance en sélection de variables sous une hypothèse d'irreprésentabilité.

Nous citons seulement les travaux récents qui concerne le cas où la dimension M est grande. Nous renvoyons à Hebiri [51] pour une présentation exhaustive des résultats existants. Bunea, Tsybakov et Wegkamp [14] établissent des inégalités oracles pour le problème de prédiction dans un modèle de régression non paramétrique sous une hypothèse de cohérence mutuelle. Zhang et Huang [114] considère le modèle de régression linéaire et établissent

une inégalité oracle pour le problème de prédiction sous une hypothèse de design incohérent plus restrictive que celle de [6]. Koltchinskii [62] et Van de Geer [99] démontrent des inégalités oracles pour les problèmes de prédiction et d'estimation avec la norme l_1 dans un cadre stochastique général avec une fonction de perte lipschitzienne sous des conditions similaire à celle de valeurs propres restreintes positives de Bickel, Ritov et Tsybakov [6]. Zhao et Yu [116] démontrent pour le modèle de régression linéaire et la perte quadratique la consistance en sélection de variables du Lasso sous une hypothèse d'irreprésentabilité forte. Meinshausen et Bühlmann [80] établissent un résultat similaire dans un cadre de modèles graphiques. Wainwright [103] démontre sous une hypothèse d'irreprésentabilité forte un résultat de sélection de variables non asymptotique. Bunea [12] démontre la consistance du Lasso sous une hypothèse de cohérence mutuelle.

Un inconvénient du Lasso est qu'il nécessite des hypothèses assez contraignantes pour établir les résultats de consistance. L'hypothèse d'irreprésentabilité couramment employée pour établir les résultats de sélection de variables dans le modèle de régression linéaire exclue le cas où les variables explicatives sont fortement corrélées, ce qui est typiquement le cas en pratique. Ceci motive l'introduction par Zou [117] du Lasso adaptatif. Cet estimateur est obtenu par minimisation du risque empirique muni de la pénalité $\lambda \sum_{j=1}^{M} W_j |\beta_j|$, où les poids W_j dépendent des données. Typiquement, ils sont de la forme $W_j = 1/|\tilde{\beta}_j|$, où $\tilde{\beta}$ est un estimateur préliminaire. Zou a montré que si l'estimateur préliminaire est consistant en estimation, alors le Lasso adaptatif sera lui aussi consistant en estimation et de plus en sélection de variables sans aucune hypothèse supplémentaire sur la matrice X que celles nécessaires pour établir la consistance de $\tilde{\beta}$. Lorsque la dimension M est fixée, le choix courant pour $\tilde{\beta}$ est l'estimateur des moindres carrés. En grande dimension, nous pouvons choisir comme estimateur préliminaire l'estimateur ridge. Néanmoins, le choix de l'estimateur préliminaire en grande dimension est une question ouverte.

Une conséquence du fait que les conditions de consistance du Lasso sont rarement satisfaites en pratique est la propension de cet estimateur à sélectionner des variables non pertinentes. Plusieurs travaux proposent des extensions de l'estimateur Lasso visant à corriger ce défaut. Ainsi, Bach [4] propose l'estimateur Bolasso pour *Bootstrapped Lasso*. Cette procédure consiste en le tirage d'un certain nombre de réplications bootstraps de l'échantillon de départ, puis du calcul de l'estimateur Lasso et de l'ensemble des variables sélectionnées pour chacun de ces échantillons bootstraps. Les variables sélectionnées finales sont obtenues par intersection des ensembles de variables sélectionnées pour chaque réplication bootstrap. Bach démontre la consistance en sélection de variables de l'estimateur Bolasso. De plus,

cet estimateur se comporte mieux en pratique puisqu'il sélectionne moins de variables non pertinentes. Meinshausen et Bülhmann [81] propose une méthode similaire basée sur la replication bootstrap de l'échantillon initial avec une randomisation supplémentaire dans la définition de la pénalité. Une autre stratégie consiste à seuiller l'estimateur Lasso. Meinshausen et Yu [82], Zhang et Huang [114] établissent la consistance d'estimation en norme l_∞ de l'estimateur Lasso sous une hypothèse de design incohérent avec des vitesses de convergence sous-optimales. Ils exploitent ensuite ce résultat pour démontrer la consistance en sélection de variables du Lasso seuillé.

Du point de vue algorithmique, le Lasso est solution d'un problème de minimisation convexe. Nous pouvons calculer une solution par programmation quadratique. Il existe d'autres algorithmes plus performants pour calculer l'estimateur Lasso qui sont réalisables même en grande dimension. Un algorithme populaire est le LARS, proposé par Efron, Hastie, Johnstone et Tibshirani [37], essentiellement parce qu'il permet d'approximer le chemin de régularisation de l'estimateur Lasso, i.e., l'ensemble des solutions Lasso $\hat{\theta}^L(r)$ lorsque le paramètre de pénalisation r varie dans $[0, \infty[$.

Le Dantzig Selector.

L'estimateur Dantzig Selector a été introduit plus récemment par Candès et Tao [16] dans un modèle de régression linéaire. Dans [16], les auteurs établissent des inégalités oracles pour la prédiction et l'estimation en norme l_2 sous une hypothèse d'isométrie restreinte sur la matrice de design. Bickel, Ritov et Tsybakov [6] établissent le lien entre le Dantzig Selector et le Lasso. Ils montrent notamment simultanément pour le Lasso et le Dantzig Selector des inégalités oracles pour la prédiction et l'estimation avec la norme l_p, $1 \leqslant p \leqslant 2$ sous une condition de valeurs propres restreintes similaire mais moins restrictive que la condition d'isométrie restreinte de [16]. Koltchinskii [63] considère le modèle de régression avec design aléatoire et prouve des résultats similaires sous une hypothèse proche de celle de [6]. Le seul résultat sur la sélection de variables pour le Dantzig Selector est [71] qui est l'objet du chapitre 1 de cette thèse.

Du point de vue algorithmique, le Dantzig Selector est solution d'un problème de minimisation convexe sous contraintes linéaires. Nous pouvons calculer une solution par programmation linéaire. James et Radchenko [55] proposent un algorithme de type LARS pour approximer le chemin de régularisation du Dantzig Selector, i.e., l'ensemble des solutions $\hat{\theta}^D(r)$ lorsque le paramètre de pénalisation r varie dans $[0, \infty[$.

Le Group Lasso.

L'estimateur Group Lasso que nous étudions dans le chapitre 4 est une version particulière adaptée au cadre multi-tâches de l'estimateur Group Lasso initialement introduit par Yuan et Lin [112]. Plusieurs articles analysent les propriétés statistiques du Group Lasso [4, 22, 54, 64, 78, 79, 83, 88]. La plupart de ces articles se concentrent sur le Group Lasso dans les modèles additifs [54, 64, 79, 88] ou bien les modèles additifs généralisés : Meier, van de Geer et Bühlmann [78]. Nardi et Rinaldo [83] établissent des inégalités d'oracles pour le Group Lasso dans un cadre général. Cependant, les bornes qu'ils obtiennent sont sous-optimales. Bach [4] démontre la consistance asymptotique en sélection de variables de l'estimateur Group Lasso. Notons aussi les travaux de Obozinski et al. [86] sur la sélection de variables avec le Group Lasso dans un modèle de régression multi-tâches plus restrictif que celui que nous considérons dans le chapitre 4 ([86] considère le cas particulier où $X_t \equiv X$ pour tout t). Dans le chapitre 4, nous mettons en évidence plusieurs avantages théoriques du Group Lasso sur le Lasso usuel pour la prédiction et l'estimation. Nous montrons notamment que l'influence de la dimension M devient négligeable pour une taille des groupes T suffisamment grande (logarithmique en la dimension M).

Du point de vue algorithmique, le Group Lasso est solution d'un problème de minimisation convexe. Liu, Ji et Ye [69] proposent une méthode de projection pour calculer une solution du problème Group Lasso avec une complexité algorithmique linéaire. Nous renvoyons à Argyriou, Evgeniou et Pontil [1] pour plus de détails sur l'aspect algorithmique.

L'estimateur à poids exponentiels moyennés.

L'estimateur à poids exponentiels moyennés est une version particulière des algorithmes de descente miroir introduits par Nemirovski et Yudin [85]. Juditsky et al. [113] établissent une inégalité oracle pour l'estimateur $\hat{\theta}^{MA}$ dans le problème de prédiction lorsque Θ est le simplexe de $\mathbb{R}^M$ avec un terme de reste optimal $\Delta(n, M, \Theta) \asymp \sqrt{(\log M)/n}$. Juditsky, Rigollet et Tsybakov [57] considèrent l'estimateur $\hat{\theta}^{MA}$ sur un ensemble Θ de cardinal fini $M \geqslant 2$ et démontrent une inégalité oracle en moyenne pour le problème de prédiction avec un terme de reste de l'ordre $\Delta(n, M) \asymp (\log M)/n$. Ces résultats sont liés aux travaux sur la prédiction on-line de suites individuelles déterministes [20, 50, 60, 102] et aussi aux avancées récentes sur la théorie PAC-Bayésienne [2, 13, 77]. Dalalyan et Tsybakov [25] établissent des inégalités oracles pour l'estimateur à poids exponentiels sans moyennage dans un modèle de régression avec design déterministe avec un terme de reste $\Delta(\theta^*, n, M) \asymp M(\theta^*)(\log M)/n$. Il n'existe pas de résultats pour l'estimation du paramètre θ^* et la sélection de variables

pour les estimateurs à poids exponentiels.

Du point de vue algorithmique, l'estimateur à poids exponentiels est extrêmement rapide à calculer lorsque l'ensemble Θ est fini. Il est en revanche plus délicat à calculer lorsque Θ est un ensemble indénombrable. Dalalyan et Tsybakov [24] montrent qu'il est possible d'approximer cet estimateur en simulant une diffusion de Langevin pour des valeurs modérément grandes de la dimension.

1.1.6 Contributions

Dans le chapitre 2, nous considérons le modèle de régression linéaire avec une matrice de design déterministe X. Notons $\Psi = X^T X/n$ la matrice de Gram associée. Nous supposons que $\Psi = (\Psi_{i,j})_{1 \leqslant i,j \leqslant M}$ vérifie une condition de cohérence mutuelle

$$\max_{i \neq j} |\Psi_{i,j}| \leqslant \frac{1}{cs}.$$

Nous établissons simultanément pour les estimateurs Lasso et Dantzig Selector une inégalité oracle pour l'estimation en norme l_∞ avec la vitesse optimale

$$\mathbb{P}\left(|\hat{\theta} - \theta^*|_\infty \leqslant Cr, \quad \forall \hat{\theta} \in \hat{\Theta}\right) \geqslant 1 - M^{1-A^2/8}, \quad r = A\sigma\sqrt{\frac{\log M}{n}}, \; A > 0$$

où C est une constante absolue et $\hat{\Theta}$ désigne soit l'ensemble des estimateurs Lasso ou Dantzig Selector. Ensuite nous prouvons la propriété de sélection de variables pour les estimateurs Lasso et Dantzig Selector seuillés sous l'hypothèse supplémentaire couramment utilisée que les composantes non nulles sont suffisamment grandes. Si $\rho = \min_{j \in J^*} |\theta^*_j| > 2Cr$, alors

$$\mathbb{P}\left(\overrightarrow{\text{sign}}(\tilde{\theta}) = \overrightarrow{\text{sign}}(\theta^*), \quad \tilde{\theta} \in \tilde{\Theta}\right) \geqslant 1 - M^{1-A^2/8},$$

où $\tilde{\Theta}$ désigne l'ensemble des estimateurs Lasso ou Dantzig Selector seuillés.

Dans le chapitre 3, nous présentons plusieurs conditions suffisantes sur la matrice de design X utilisées dans la littérature sur les propriétés statistiques des estimateurs Lasso et Dantzig Selector. Donoho et Tanner [31] ont établi la condition nécessaire et suffisante de reconstruction parfaite du vector θ^* par minimisation l_1 sous contrainte dans un modèle de régression linéaire sans bruit. Juditsky et Nemirovski [56] prouvent un résultat similaire et proposent des conditions suffisantes vérifiables en pratique. Nous étudions les relations entre ces différentes conditions. Nous proposons une preuve directe du résultat sur la condition nécessaire et suffisante de reconstruction parfaite par minimisation l_1 de Juditsky et Nemirovski [56]. En présence de bruit aléatoire, nous établissons une inégalité oracle pour

l'estimation en norme l_1 pour le Dantzig Selector sous l'hypothèse de [31, 56] et pour le Lasso sous une hypothèse plus restrictive.

Dans le chapitre 4, nous considérons le même modèle de régression linéaire avec design déterministe. Nous montrons sous une condition d'irreprésentabilité sur la matrice de design que le problème de minimisation Lasso admet une unique solution avec probabilité proche de 1 et nous donnons sa forme explicite. Nous établissons une inégalité oracle pour l'estimation l_∞ et la sélection de variables. Prenons $r = A\sigma\sqrt{(\log M)/n}$ avec $A > 2\sqrt{2}$. Supposons que le bruit est gaussien. Si les composantes non nulles du vecteur cible θ^* sont suffisamment grandes (cf. l'hypothèse 4.3) et si $M(\theta^*) \leqslant s$, alors nous avons, avec probabilité au moins $1 - 2M^{1-\frac{(\eta A)^2}{2}} - sM^{-\frac{A^2}{2}}$ où $0 < \eta < 1/2$, que

- la solution Lasso $\hat{\theta}^L$ est unique et égale à $\tilde{\theta}^0$ défini dans (4.6)-(4.7).
- De plus, nous avons

$$|\hat{\theta}^L - \theta^*|_\infty \leqslant rd^*,$$

où d^* est défini dans (4.15) et

$$\overrightarrow{\text{sign}}(\hat{\theta}^L) = \overrightarrow{\text{sign}}(\theta^*).$$

Nous discutons de l'optimalité de la borne d'estimation obtenue sous cette hypothèse plus faible.

Dans le chapitre 5, nous étudions les propriétés statistiques du Group Lasso dans un modèle de régression multi-tâches. Sous des hypothèses similaires à celles utilisées dans [6, 71] pour le Lasso et le Dantzig Selector, nous établissons que, avec probabilité au moins $1 - M^{1-q}$ où $q = \min\{8\log M, A\sqrt{T}/8\}$,

$$\begin{aligned} R(\hat{\theta}^{GL}) - R(\theta^*) &\leqslant C\sigma\frac{s}{n}\left(1 + \frac{A\log M}{\sqrt{T}}\right) \\ \frac{1}{\sqrt{T}}\sum_{j=1}^{M}\|(\hat{\theta}^{GL} - \theta^*)^j\| &\leqslant \frac{C}{2}\sigma\frac{s}{\sqrt{n}}\sqrt{1 + \frac{A\log M}{\sqrt{T}}}, \end{aligned}$$

où $C > 0$ est une constante qui dépend de la condition sur la matrice de design. Nous constatons que pour un nombre de tâches $T \geqslant (\log M)^2$, l'effet de la grande dimension M est annulé. Le Lasso non groupé ne jouit pas d'une telle propriété. Nous établissons des résultats pour l'estimation en norme l_∞ et la sélection de variables sous une hypothèse de cohérence mutuelle. Pour traiter le cas où le bruit n'est pas gaussien, nous démontrons notamment une généralisation de l'inégalité de Nemirovski qui présente un intérêt propre.

Dans le Chapitre 6, nous considérons un problème d'optimisation stochastique général. Nous étudions une généralisation de l'estimateur Dantzig Selector dans ce cadre. Nous établissons pour ce nouvel estimateur des inégalités oracles en prédiction et en estimation pour la norme l_1 de type (1.1) et (1.2). Nous considérons des applications à la régression logistique et la régression linéaire avec une perte lipschitzienne. Pour ce dernier modèle, nous démontrons aussi un résultat d'estimation en norme l_∞ et un résultat sur la sélection de variables pour l'estimateur Dantzig généralisé et Dantzig usuel.

Dans le chapitre 7, nous considérons l'estimateur à poids exponentiels moyenné $\hat{\theta}^{MA}$ dans notre cadre d'optimisation stochastique général. Les performances de l'estimateur $\hat{\theta}^{MA}$ dépendent de manière cruciale de la loi a priori Π utilisée pour le calculer. Nous proposons plusieurs choix possibles pour Π et nous établissons notamment des inégalités oracles pour la prédiction de la forme

$$\mathbb{E}(R(\hat{\theta}^{MA})) \leqslant \min_{\theta\in\Theta}\left\{R(\theta) + C\frac{M(\theta)}{n}\log\left(\frac{Mn}{eM(\theta)}\right)\right\},$$

où Θ est une boule l_1 de $\mathbb{R}^M$.

1.2 Aggrégation d'estimateurs de densité pour la norme L^π, $1 \leqslant \pi \leqslant \infty$

Nous nous intéressons à l'agrégation d'estimateurs dans le cadre de l'estimation d'une densité de probabilité. Les inégalités oracles existantes concernent majoritairement la perte L^2 par des procédures de poids exponentiels dans Juditsky, Rigollet et Tsybakov [57] ou bien par minimisation du risque empirique dans Samarov et Tsybakov [92]. Il existe aussi des résultats pour la perte L^1 : Birgé [7] et Devroye et Lugosi [28] mais leurs procédures d'agrégation sont difficilement réalisables en pratique.

Nous proposons d'adapter une procédure d'agrégation initialement proposée par Goldenshluger [46] pour le modèle de bruit blanc au modèle de densité. Nous établissons des inégalités oracle pour la norme L^π, $1 \leqslant \pi \leqslant \infty$. Puis nous exploitons cette procédure pour construire un estimateur de la densité adaptatif en la régularité et minimax pour la norme L^π avec $1 \leqslant \pi \leqslant 2$.

Nous adaptons ensuite la procédure de Goldenshluger et Lepski [47] à notre cadre d'estimation de densité pour construire une procédure d'estimation minimax et adaptative en la régularité sur une classe de Besov pour la norme L^∞.

1.2.1 Agrégation pour la norme L^π

Soit $X_1, \ldots, X_n$ des variables aléatoires i.i.d. à valeurs dans $\mathbb{R}^d$ de densité $f \in L^\pi(\mathbb{R}^d)$, avec $1 \leqslant \pi \leqslant \infty$. Pour une fonction $g \in L^\pi(\mathbb{R}^d)$ et $\pi < \infty$, nous définissons $||g||_\pi = \left(\int_{\mathbb{R}^d} |g(x)|^\pi dx\right)^{1/\pi}$. Si $\pi = \infty$, nous posons $||g||_\infty = \operatorname{ess\,sup}_{x\in\mathbb{R}^d} |g(x)|$.

La performance d'un estimateur $\hat{f}$ construit à partir de $\mathbb{X}_n = (X_1, \ldots, X_n)$ est mesurée par le risque L^π :

$$R_{n,\pi}(\hat{f}, f) = E_f^{\otimes n} \|\hat{f} - f\|_\pi^\pi,$$

où $E_f^{\otimes n}$ désigne l'espérance par rapport à la loi $P_f^{\otimes n}$ de l'échantillon $\mathbb{X}_n$.

Soit une famille de fonctions $\mathcal{F}_M = \{f_1, \ldots, f_M\}$ sur $\mathbb{R}^d$. Notons que les fonctions de l'ensemble $\mathcal{F}_M$ peuvent être aussi bien des fonctions déterministes que des estimateurs de la densité f construits à partir d'un échantillon préliminaire. Le but de notre procédure d'agrégation est de contruire un estimateur $\hat{f}$ tel que

$$R_{n,\pi}(\hat{f}, f) \leqslant C(\pi) \min_{1\leqslant j\leqslant M} R_{n,\pi}(f_j, f) + \Delta(n, \mathcal{F}_M),$$

où $C(\pi) \geqslant 1$ et $\Delta(n, \mathcal{F}_M)$ est un terme de reste qui ne dépend pas de f. Le terme $\min_{1\leqslant j\leqslant M} R_{n,\pi}(f_j, f)$ est appelé risque oracle. Désormais, nous appelerons agrégat l'estimateur agrégé $\hat{f}$.

1.2.2 Estimation adaptative

L'inégalité d'oracle ci-dessus est un résultat non asymptotique qui peut être exploité pour construire des estimateurs adaptatifs au sens minimax. Considérons le cas où $\mathcal{F}_M$ est une famille d'estimateurs. Selon la densité f à estimer, un estimateur donné f_j aura des performances d'estimation plus ou moins bonnes. Néanmoins, si pour toute densité f dans une classe $\mathcal{F}$, il existe un estimateur $f_j \in \mathcal{F}_M$ qui possède de bonnes performances pour l'estimation de cette densité particulière f, alors l'agrégat $\hat{f}$ sera performant uniformément sur la classe $\mathcal{F}$ si le terme de reste $\Delta(n, \mathcal{F}_M)$ est suffisamment petit.

Nous précisons maintenant la notion de performance d'un estimateur selon le critère minimax. Soit $\mathcal{F}$ une classe de densités de probabilité.

Definition 1.1. *Une suite positive (v_n) est appelée vitesse de convergence minimax sur la classe $\mathcal{F}$ si*

– *Il existe une constante $c > 0$ telle que*

$$\liminf_{n\to\infty} v_n^{-1} [\inf_{T_n} \sup_{f\in\mathcal{F}} R_{n,\pi}(T_n, f)] \geqslant c, \tag{1.9}$$

où $\inf_{T_n}$ *désigne l'infimum sur l'ensemble de tous les estimateurs.*

– *Il existe une constante* $C > 0$ *et un estimateur* $\hat{f}_n$ *tels que*

$$\limsup_{n\to\infty} v_n^{-1}[\sup_{f\in\mathcal{F}} R_{n,\pi}(\hat{f}_n, f)] \leqslant C. \tag{1.10}$$

Un estimateur vérifiant (1.10) lorsque (1.9) est vérifié est appelé estimateur optimal en vitesse de convergence.

Notons que la vitesse minimax (v_n) est définie à une constante multiplicative près. Précisons de plus que cette vitesse dépend à la fois de la classe de densités étudiée $\mathcal{F}$ et du critère de risque choisi.

Supposons maintenant que la classe de densité $\mathcal{F}$ inconnue appartient à une famille de classes $\{\mathcal{F}_\gamma\}_{\gamma\in\Gamma}$.

Definition 1.2. *Un estimateur* $\hat{f}_n$ *est dit adaptatif en vitesse de convergence sur la famille* $\{\mathcal{F}_\gamma\}_{\gamma\in\Gamma}$ *s'il existe une constante* $C > 0$ *telle que*

$$\limsup_{n\to\infty} \sup_{\gamma\in\Gamma} [v_n^{-1}(\gamma) \sup_{f\in\mathcal{F}_\gamma} R_{n,\pi}(\hat{f}_n, f)] \leqslant C.$$

Nous proposons dans le chapitre 7 d'agréger des estimateurs par ondelettes avec seuillage [26] pour construire une procédure adaptative en la vitesse de convergence quand la densité f appartient à un espace de Besov.

1.2.3 Bibliographie

Le problème d'agrégation d'estimateurs de densité a été étudié dans [8, 18, 57, 92, 109, 115] avec la divergence de Kullback-Leibler et le risque L^2. Devroye et Lugosi [28] et Birgé [7] ont obtenu des résultats pour le risque L^1. L'optimalité des vitesses d'agrégation au sens de [97] a été établie dans Rigollet et Tsybakov [89] pour le risque L^2 et dans Lecué [65] pour la divergence de Kullback-Leibler et le risque L^1.

La littérature sur le problème d'estimation de densité par ondelettes est vaste. Nous ne prétendons pas être exhaustif sur le sujet. Donoho, Johnstone, Kerkyacharian et Picard [32, 33] établissent les vitesses minimax d'estimation sur les classes de Besov et proposent une procédure d'estimation par ondelettes avec seuillage minimax et adaptative en la régularité à un terme logarithmique près. Delyon et Juditsky [26] proposent un estimateur par ondelettes avec seuillage minimax mais non adaptatif en la régularité. Chesneau et Lecué [23] proposent une procédure minimax et adaptive en la régularité pour la perte L^2 basée sur l'agrégation à poids exponentiels des estimateurs à ondelettes seuillés de [26].

1.2.4 Contributions

Dans le chapitre 8, nous étudions deux procédures d'agrégation dans un modèle de densité. Pour la première procédure, $\mathcal{F}_M$ est constitué de fonctions déterministes. Nous établissons l'inégalité oracle intégrée suivante pour l'agrégat $\hat{f}$ construit à partir de l'échantillon $\mathbb{X}_n$:

$$E_f^{\otimes n}\left(||\hat{f}-f||_\pi^\alpha\right) \leqslant C_1 \min_{1\leqslant j\leqslant M} ||f_j - f||_\pi^\alpha + C_2\left(Q_2(\pi)\frac{\log M}{n}\right)^\alpha + C_3\left(Q_1(\pi)\sqrt{||f||_\infty \frac{\log M}{n}}\right)^\alpha, \tag{1.11}$$

où $\alpha > 0$,

$$Q_1(\pi) = \max_{i\neq j} \frac{||f_j - f_i||_{2\pi-2}^{\pi-1}}{||f_j - f_i||_\pi^{\pi-1}}, \tag{1.12}$$

$$Q_2(\pi) = \max_{i\neq j} \frac{||f_j - f_i||_\infty^{\pi-1}}{||f_j - f_i||_\pi^{\pi-1}}, \tag{1.13}$$

et les constantes C_j, $1 \leqslant j \leqslant 3$ sont explicitées. Nous montrons que les vitesses d'agrégation obtenues sont optimales en établissant les bornes inférieures. Ainsi les termes $Q_j(\pi)$, $j = 1, 2$, qui dépendent explicitement du dictionnaire, ne peuvent être supprimés. Par conséquent, ces procédures sont difficiles à exploiter pour construire des estimateurs adaptatifs puisque cela nécessite l'étude délicate des ratios $Q_j(\pi)$, $j = 1, 2$. Nous pouvons néanmoins pour certaines familles d'estimateurs, notamment les estimateurs à ondelettes, démontrer des bornes supérieures sur les termes $Q_j(\pi)$ et aboutir à des résultats d'adaptivité.

Nous adaptons la procédure de Goldenshluger et Lepski [47] au problème d'estimation de densité en norme L^∞. Dans ce cas, $\mathcal{F}_M = \{\hat{f}_1, \ldots, \hat{f}_M\}$ est constitué d'estimateurs linéaires construit à partir de l'échantillon $\mathbb{X}_n$. Soit $\hat{f}$ l'agrégat construit par cette procédure à partir du même échantillon $\mathbb{X}_n$. Nous obtenons l'inégalité oracle suivante :

$$E_f^{\otimes n}\|\hat{f}-f\|_\infty \leqslant C(\kappa)\left(E_f^{\otimes n}(\min_{1\leqslant j\leqslant M}\|\hat{f}_j - f\|_\infty) + P_f^{\otimes n}(\mathcal{A}_\kappa^c)\right),$$

où $C(\kappa) \geqslant 1$ et $\mathcal{A}_\kappa$ est un évènement de probabilité proche de 1.

Nous exploitons ensuite ces procédures d'agrégation pour construire des estimateurs minimax et adaptifs en la régularité lorsque la densité inconnue f appartient à une boule d'un espace de Besov. Dans un premier temps, nous considérons comme dictionnaire les estimateurs minimax et non adaptatifs de Delyon et Juditsky [26]. A partir de ce dictionnaire et de la première procédure d'agrégation, nous construisons un estimateur minimax et adaptatif

en la régularité pour la norme L^π avec $1 \leqslant \pi \leqslant 2$. Le cas $\pi > 2$ est en cours d'investigation. Puis nous considérons les estimateurs à ondelettes linéaires et la seconde procédure et nous montrons un résultat similaire pour la norme L^∞.

Avertissement : Chaque chapitre de cette partie se présente avec ses propres notations et peut être lu indépendamment des autres. Néanmoins, la plupart des notations sont communes à tous les chapitres.

foreword :

Each chapter of this thesis has its own system of notations and can be read independently of the others. However, most of the notations is common to all the chapters.

Chapter 2

Sup-norm convergence rate and sign concentration property of the Lasso and the Dantzig Selector

We derive the l_∞ convergence rate simultaneously for Lasso and Dantzig estimators in a high-dimensional linear regression model under a mutual coherence assumption on the Gram matrix of the design and two different assumptions on the noise: Gaussian noise and general noise with finite variance. Then we prove that the thresholded Lasso and Dantzig estimators with a proper choice of the threshold simultaneously enjoy a sign concentration property provided that the non-zero components of the target vector are not too small.

2.1 Introduction

The Lasso is an l_1 penalized least squares estimator in linear regression models proposed by Tibshirani [94]. The Lasso enjoys two important properties. First, it is sparse by construction, i.e., it has a large number of zero components. Second, it is computationally feasible even for high-dimensional data (Efron et al. [37], Osborne et al. [87]) whereas classical procedures such as BIC are not feasible when the number of parameters becomes large. The first property raises the question of model selection consistency of Lasso, i.e., of identification of the subset of non-zero parameters. A closely related problem is the one of sign consistency, or differently put, the problem of identification of the non-zero parameters and their signs (cf. Bach [3], Bunea [12], Meinshausen and Bühlmann [80], Meinshausen and Yu [82], Wainwright [104], Zhao and Yu [116] and the references cited therein).

Zou [117] has proved estimation and variable selection results for the adaptive Lasso: a variant of the Lasso where the weights on the different components in the l_1 penalty vary and are made data dependent. We also mention the works on the convergence of the Lasso estimator under the prediction loss: Bickel, Ritov and Tsybakov [6], Bunea, Tsybakov and Wegkamp [14], Greenshtein and Ritov [48], Koltchinskii [62, 63], Van der Geer [99, 100].

Knight and Fu [61] have proved the estimation consistency of the Lasso estimator in the case where the number of parameters is fixed and smaller than the sample size. The l_2 consistency of Lasso with convergence rate has been proved in Bickel, Ritov and Tsybakov [6], Meinshausen and Yu [82], Zhang and Huang [114]. These results trivially imply the l_p consistency, with $2 \leqslant p \leqslant \infty$, however with a suboptimal rate (cf., e.g., Theorem 3 in [114]). Bickel, Ritov and Tsybakov [6] have proved that the Dantzig selector of Candes and Tao [16] shares a lot of common properties with the Lasso. In particular they have shown simultaneous l_p consistency with rates of the Lasso and Dantzig estimators for $1 \leqslant p \leqslant 2$. To our knowledge, there is no result on the l_∞ convergence rate and sign consistency of the Dantzig estimator.

The notion of l_∞ and sign consistency should be properly defined when the number of parameters is larger than the sample size. We may have indeed an infinity of possible target vectors and solutions to the Lasso and Dantzig minimization problems. This difficulty is not discussed in [12, 80, 82, 104, 114] where either the target vector or the Lasso estimator or both are assumed to be unique. We show that under a sparsity scenario, it is possible to derive l_∞ and sign consistency results even when the number of parameters is larger than the sample size. We refer to Theorem 6.3 and the Remark 1, p. 21, in [6] which suggest a

way to clarify the difficulty mentioned above.

In this chapter, we consider a high-dimensional linear regression model where the number of parameters can be much greater than the sample size. We show that under a mutual coherence assumption on the Gram matrix of the design, the target vector which has few non-zero components is unique. We do not assume the Lasso or Dantzig estimators to be unique. We establish the l_∞ convergence rate of all the Lasso and Dantzig estimators simultaneously under two different assumptions on the noise. The rate that we get improves upon those obtained for the Lasso in the previous works. Then we show a sign concentration property of all the thresholded Lasso and Dantzig estimators simultaneously for a proper choice of the threshold if we assume that the non-zero components of the sparse target vector are large enough. Our condition on the size of the non-zero components of the target vector is less restrictive than in [104, 114, 116]. In addition, we prove analogous results for the Dantzig estimator, which to our knowledge was not done before.

The chapter is organized as follows. In Section 2.2, we present the Gaussian linear regression model, the assumptions, the results and we compare them with the existing results in the literature. In Section 2.3, we consider a general noise with zero mean and finite variance and we show that the results remain essentially the same, up to a slight modification of the convergence rate. In Section 2.4, we provide the proofs of the results.

2.2 Model and Results

Consider the linear regression model

$$Y = X\theta^* + W, \tag{2.1}$$

where X is an $n \times M$ deterministic matrix, $\theta^* \in \mathbb{R}^M$ and $W = (W_1, \dots, W_n)^T$ is a zero-mean random vector such that $\mathbb{E}[W_i^2] \leqslant \sigma^2$, $1 \leqslant i \leqslant n$ for some $\sigma^2 > 0$. For any $\theta \in \mathbb{R}^M$, define $J(\theta) = \{j \ : \ \theta_j \neq 0\}$. Let $M(\theta) = |J(\theta)|$ be the cardinality of $J(\theta)$ and $\vec{\text{sign}}(\theta) = (\text{sign}(\theta_1), \dots, \text{sign}(\theta_M))^T$ where

$$\text{sign}(t) = \begin{cases} 1 & \text{if } t > 0, \\ 0 & \text{if } t = 0, \\ -1 & \text{if } t < 0. \end{cases}$$

For any vector $\theta \in \mathbb{R}^M$ and any subset J of $\{1, \dots, M\}$, we denote by θ_J the vector in $\mathbb{R}^M$ which has the same coordinates as θ on J and zero coordinates on the complement J^c

of J. For any integers $1 \leqslant d, p < \infty$ and $z = (z_1, \ldots, z_d) \in \mathbb{R}^d$, the l_p norm of the vector z is denoted by $|z|_p \triangleq \left(\sum_{j=1}^d |z_j|^p\right)^{1/p}$, and $|z|_\infty \triangleq \max_{1 \leqslant j \leqslant d} |z_j|$.

Note that the assumption of uniqueness of θ^* is not satisfied if $M > n$. In this case, if a vector $\theta^* = \theta^0$ satisfies (2.1), then there exists an affine space $\Theta^* = \{\theta^* : X\theta^* = X\theta^0\}$ of dimension larger than $M - n$ of vectors satisfying (2.1). So the question of sign consistency becomes a problem when $M > n$ because we can easily find two distinct vectors θ and θ' satisfying (2.1) such that $\vec{\text{sign}}(\theta) \neq \vec{\text{sign}}(\theta')$. However we will show that under our assumptions, the vector θ^* is unique.

The Lasso and Dantzig estimators $\hat{\theta}^L$, $\hat{\theta}^D$ solve respectively the minimization problems

$$\min_{\theta \in \mathbb{R}^M} \frac{1}{n}|Y - X\theta|_2^2 + 2r|\theta|_1, \tag{2.2}$$

and

$$\min_{\theta \in \mathbb{R}^M} |\theta|_1 \text{ subject to } \left|\frac{1}{n}X^T(Y - X\theta)\right|_\infty \leqslant r, \tag{2.3}$$

where $r > 0$ is a constant. A convenient choice in our context will be $r = A\sigma\sqrt{(\log M)/n}$, for some $A > 0$. We denote respectively by $\hat{\Theta}^L$ and $\hat{\Theta}^D$ the set of solutions to the Lasso and Dantzig minimization problems (2.2) and (2.3).

The definition of the Lasso minimization problem we use here is not the same as the one in [94], where it is defined as

$$\min_{\theta \in \mathbb{R}^M} \frac{1}{n}|Y - X\theta|_2^2 \text{ subject to } |\theta|_1 \leqslant t,$$

for some $t > 0$. However these minimization problems are strongly related, cf. [21]. The Dantzig estimator was introduced and studied in [16]. Define $\Phi(\theta) = \frac{1}{n}|Y - X\theta|_2^2 + 2r|\theta|_1$. A necessary and sufficient condition for a vector θ to minimize Φ is that the zero vector in $\mathbb{R}^M$ belongs to the subdifferential of Φ at point θ, i.e.,

$$\begin{cases} \frac{1}{n}(X^T(Y - X\theta))_j = \text{sign}(\theta_j)r & \text{if } \theta_j \neq 0, \\ \left|\frac{1}{n}(X^T(Y - X\theta))_j\right| \leqslant r & \text{if } \theta_j = 0. \end{cases}$$

Thus, any vector $\theta \in \hat{\Theta}^L$ satisfies the Dantzig constraint

$$\left|\frac{1}{n}X^T(Y - X\theta)\right|_\infty \leqslant r. \tag{2.4}$$

The Lasso estimator is unique if $M < n$, since in this case $\Phi(\theta)$ is strongly convex. However, for $M > n$ it is not necessarily unique. The uniqueness of Dantzig estimator is not granted either. From now on, we set $\hat{\Theta} = \hat{\Theta}^L$ or $\hat{\Theta}^D$ and $\hat{\theta}$ denotes an element of $\hat{\Theta}$.

Now we state the assumptions on our model. The first assumption concerns the noise variables.

Assumption 2.1. *The random variables $W_1, \ldots, W_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$.*

We also need assumptions on the Gram matrix

$$\Psi \triangleq \frac{1}{n} X^T X.$$

Assumption 2.2. *The elements $\Psi_{i,j}$ of the Gram matrix Ψ satisfy*

$$\Psi_{j,j} = 1, \ \forall 1 \leqslant j \leqslant M, \tag{2.5}$$

and

$$\max_{i \neq j} |\Psi_{i,j}| \leqslant \frac{1}{\alpha(1 + 2c_0)s}, \tag{2.6}$$

for some integer $s \geqslant 1$ and some constant $\alpha > 1$, where $c_0 = 1$ if we consider the Dantzig estimator, and $c_0 = 3$ if we consider the Lasso estimator.

The notion of mutual coherence was introduced in [34] where the authors required that $\max_{i \neq j} |\Psi_{i,j}|$ were sufficiently small. Assumption 2.2 is stated in a slightly weaker form in [6]-[15].

Consider two vectors θ^1 and θ^2 satisfying (2.1) such that $M(\theta^1) \leqslant s$ and $M(\theta^2) \leqslant s$. Denote $\theta = \theta^1 - \theta^2$ and $J = J(\theta^1) \cup J(\theta^2)$. We clearly have $X\theta = 0$ and $|J| \leqslant 2s$. Assume that $\theta \neq 0$. Under Assumption 2.2, similarly as we derive the inequality (2.11) in Section 2.4 below and using the fact that $|\theta|_1 \leqslant \sqrt{2s}|\theta|_2$, we get that

$$\frac{|X\theta|_2^2}{n|\theta|_2^2} > 0.$$

This contradicts the fact that $X\theta = 0$. Thus we have $\theta^1 = \theta^2$. We have proved that under Assumption 2.2 the vector θ^* satisfying (2.1) with $M(\theta^*) \leqslant s$ is unique.

Our first result concerns the l_∞ rate of convergence of Lasso and Dantzig estimators.

Theorem 2.1. *Take $r = A\sigma\sqrt{(\log M)/n}$ and $A > 2\sqrt{2}$. Let Assumptions 2.1,2.2 be satisfied. If $M(\theta^*) \leqslant s$, then*

$$\mathbb{P}\left(\sup_{\hat{\theta} \in \hat{\Theta}} \left|\hat{\theta} - \theta^*\right|_\infty \leqslant c_2 r\right) \geqslant 1 - M^{1 - A^2/8},$$

with $c_2 = \frac{3}{2}\left(1 + \frac{(1+c_0)^2}{(1+2c_0)(\alpha-1)}\right)$.

Theorem 2.1 states that in high dimensions (M large), the set of estimators $\hat{\Theta}$ is necessarily well concentrated around the vector θ^*. A similar phenomenon was already observed in [6], cf. Remark 1, page 21, for concentration in l_p norms, $1 \leqslant p \leqslant 2$. Note that c_2 in Theorem 2.1 is an absolute constant. Using Theorem 2.1, we can easily prove the consistency of the Lasso and Dantzig estimators simultaneously when $n \to \infty$. We allow the quantities s, M, $\hat{\Theta}$, θ^* to vary with n. In particular, we assume that

$$M \to \infty \text{ and } \lim_{n\to\infty} \frac{\log M}{n} = 0,$$

as $n \to \infty$, and that Assumptions 2.1,2.2 hold true for any n. Then we have

$$\sup_{\hat{\theta}\in\hat{\Theta}} \left|\hat{\theta} - \theta^*\right|_\infty \to 0 \tag{2.7}$$

in probability, as $n \to \infty$. The condition $(\log M)/n \to 0$ means that the number of parameters cannot grow arbitrarily fast when $n \to \infty$. We have the restriction $M = o(\exp(n))$, which is natural in this context.

A result on l_∞ consistency of Lasso has been previously stated in Theorem 3 of [114], where $\hat{\theta}^L$ was assumed to be unique and under another assumption on the matrix Ψ. It is not directly related to our Assumption 2.2, but can be deduced from a restricted version of Assumption 2.2 where α is taken to be substantially larger than 1. The result in [114] is a trivial consequence of the l_2 consistency, and has therefore the rate $|\hat{\theta}^L - \theta^*|_\infty = O_{\mathbb{P}}(s^{1/2}r)$ which is slower than the correct rate given in Theorem 2.1. In fact, the rate in [114] depends on the unknown sparsity s which is not the case in Theorem 2.1. Note also that Theorem 3 in [114] concerns the Lasso only, whereas our result covers simultaneously the Lasso and Dantzig estimators.

We now study the sign consistency. We make the following assumption.

Assumption 2.3. *There exists an absolute constant $c_1 > 0$ such that*

$$\rho \stackrel{\Delta}{=} \min_{j\in J(\theta^*)} |\theta_j^*| > c_1 r.$$

We will take $r = A\sigma\sqrt{(\log M)/n}$. We can find similar assumptions on ρ in the work on sign consistency of the Lasso estimator mentioned above. More precisely, the lower bound on ρ is of the order $s^{1/4}r^{1/2}$ in [82], $n^{-\delta/2}$ with $0 < \delta < 1$ in [104, 116], $\sqrt{(\log Mn)/n}$ in [12] and $\sqrt{s}r$ in [114]. Note that our assumption is the less restrictive.

We now introduce the thresholded Lasso and Dantzig estimators. For any $\hat{\theta} \in \hat{\Theta}$ the associated thresholded estimator $\tilde{\theta} \in \mathbb{R}^M$ is defined by

$$\tilde{\theta}_j = \begin{cases} \hat{\theta}_j, & \text{if } |\hat{\theta}_j| > c_2 r, \\ 0 & \text{elsewhere.} \end{cases}$$

Denote by $\tilde{\Theta}$ the set of all such $\tilde{\theta}$. We have first the following non-asymptotic result that we call sign concentration property.

Theorem 2.2. *Take $r = A\sigma\sqrt{(\log M)/n}$ and $A > 2\sqrt{2}$. Let Assumptions 2.1-2.3 be satisfied. We assume furthermore that $c_1 > 2c_2$, where c_2 is defined in Theorem 2.1. Then*

$$\mathbb{P}\left(\vec{\text{sign}}(\tilde{\theta}) = \vec{\text{sign}}(\theta^*),\ \forall \tilde{\theta} \in \tilde{\Theta}\right) \geqslant 1 - M^{1-A^2/8}.$$

Theorem 2.2 guarantees that every vector $\tilde{\theta} \in \tilde{\Theta}$ and θ^* share the same signs with high probability. Letting n and M tend to ∞ we can deduce from Theorem 2.2 an asymptotic result under the following additional assumption.

Assumption 2.4. *We have $M \to \infty$ and $\lim_{n\to\infty} \frac{\log M}{n} = 0$, as $n \to \infty$.*

Then the following asymptotic result called sign consistency follows immediately from Theorem 2.2.

Corollary 2.1. *Let the assumptions of Theorem 2.2 hold for any n large enough. Let Assumption 2.4 be satisfied. Then*

$$\mathbb{P}\left(\vec{\text{sign}}(\tilde{\theta}) = \vec{\text{sign}}(\theta^*),\ \ \forall \tilde{\theta} \in \tilde{\Theta}\right) \to 1,$$

as $n \to \infty$.

The sign consistency of Lasso was proved in [80, 116] with the Strong Irrepresentable Condition on the matrix Ψ which is somewhat different from ours. Papers [80, 116] assume a lower bound on ρ of the order $n^{-\delta/2}$ with $0 < \delta < 1$, whereas our Assumption 2.3 is less restrictive. Note also that these works assume $\hat{\theta}^L$ to be unique. Wainwright [104] does not assume $\hat{\theta}^L$ to be unique and discusses sign consistency of Lasso under a mutual coherence assumption on the matrix Ψ and the following condition on the lower bound: $\sqrt{(\log M)/n} = o(\rho)$ as $n \to \infty$, which is more restrictive than our Assumption 2.3. In particular Proposition 1 in [104] states that as $n \to \infty$, if the sequence of θ^* satisfies the above condition for all n large enough, then

$$\mathbb{P}\left(\exists \hat{\theta}^L \in \hat{\Theta}^L \text{ s.t. } \vec{\text{sign}}(\hat{\theta}^L) = \vec{\text{sign}}(\theta^*)\right) \to 1.$$

This result does not guarantee sign consistency for all the estimators $\hat{\theta}^L \in \hat{\Theta}^L$ but only for some unspecified subsequence that is not necessarily the one chosen in practice. On the contrary, Corollary 2.1 guarantees that all the thresholded Lasso and Dantzig estimators and

θ^* share the same sign vector asymptotically. It follows from this result that any solution selected by the minimization algorithm is covered and that the case $M > n$, where the set $\hat{\Theta}$ is not necessarily reduced to an unique estimator, can still be treated. We note also that the articles mentioned above treat the sign consistency for the Lasso only, whereas we prove it simultaneously for Lasso and Dantzig estimators. An improvement in the conditions that we get is probably due to the fact that we consider thresholded Lasso and Dantzig estimators. In addition note that not only the consistency results, but also the exact non-asymptotic bounds are provided by Theorems 2.1 and 2.2.

2.3 Convergence rate and sign consistency under a general noise

In the literature on Lasso and Dantzig estimators, the noise is usually assumed to be Gaussian [6, 16, 80, 104, 114] or admitting a finite exponential moment [12, 82]. The exception is the paper by Zhao and Yu [116] who proved the sign consistency of the Lasso when the noise admits a finite moment of order $2k$ where $k \geqslant 1$ is an integer. An interesting question is to determine whether the results of the previous section remain valid under less restrictive assumption on the noise. In this section, we only assume that the random variables $W_i, i = 1, \ldots, n$, are independent with zero mean and finite variance $\mathbb{E}[W_i^2] \leqslant \sigma^2$. We show that the results remain similar. We need the following assumption

Assumption 2.5. *The matrix X is such that*

$$\frac{1}{n}\sum_{i=1}^{n} \max_{1 \leqslant j \leqslant M} |X_{i,j}|^2 \leqslant c',$$

for a constant $c' > 0$.

For example, if all $X_{i,j}$ are bounded in absolute value by a constant uniformly in i, j, then Assumption 4 is satisfied. The next theorem gives the l_∞ rate of convergence of Lasso and Dantzig estimators under a mild noise assumption.

Theorem 2.3. *Assume that W_i are independent random variables with $\mathbb{E}[W_i] = 0$, $\mathbb{E}[W_i^2] \leqslant \sigma^2$, $i = 1, \ldots, n$. Take $r = \sigma\sqrt{\frac{(\log M)^{1+\delta}}{n}}$, with $\delta > 0$. Let Assumptions 2.2-2.5 be satisfied. Then*

$$\mathbb{P}\left(\sup_{\hat{\theta} \in \hat{\Theta}} \left|\hat{\theta} - \theta^*\right|_\infty \leqslant c_2 r\right) \geqslant 1 - \frac{c}{(\log M)^\delta},$$

where c_2 is defined in Theorem 2.1, and $c > 0$ is a constant depending only on c'.

Therefore the l_∞ convergence rate under the bounded second moment noise assumption is only slightly slower than the one obtained under the Gaussian noise assumption and the concentration phenomenon is less pronounced. If we assume that $\lim_{n\to\infty}(\log M)^{1+\delta}/n = 0$ and that Assumptions 2.2,2.3 and 2.5 hold true for any n with $r = \sigma\sqrt{(\log M)^{1+\delta}/n}$, then the sign consistency of thresholded Lasso and Dantzig estimators follows from our Theorem 2.3 similarly as we have proved Theorem 2.2 and Corollary 2.1. Zhao and Yu [116] stated in their Theorem 3 a result on the sign consistency of Lasso under the finite variance assumption on the noise. They assumed $\hat{\theta}^L$ to be unique and the matrix X to satisfy the condition $\max_{1\leqslant i\leqslant n}(\sum_{j=1}^M X_{i,j}^2)/n \to 0$, as $n \to \infty$. This condition is rather strong. It does not hold if $M > n$ and all the $X_{i,j}$ are bounded in absolute value by a constant. In addition, [116] assumes that the dimension $M = O(n^\delta)$ with $0 < \delta < 1$, whereas we only need that $M = o(\exp(n^{1/(1+\delta)}))$ with $\delta > 0$. Note also that [116] proves the sign consistency for the Lasso only, whereas we prove it for thresholded Lasso and Dantzig estimators.

2.4 Proofs

We begin by stating and proving two preliminary lemmas. The first lemma originates from Lemma 1 of [14] and Lemma 2 of [6].

Lemma 2.1. *Let Assumption 2.1 and (2.5) of Assumption 2.2 be satisfied. Take* $r = A\sigma\sqrt{(\log M)/n}$. *Here* $\hat{\Theta}$ *denotes either* $\hat{\Theta}^L$ *or* $\hat{\Theta}^D$. *Then we have, on an event of probability at least* $1 - M^{-A^2/8}$, *that*

$$\sup_{\hat{\theta}\in\hat{\Theta}} \left|\Psi(\theta^* - \hat{\theta})\right|_\infty \leqslant \frac{3r}{2}, \tag{2.8}$$

and for all $\hat{\theta} \in \hat{\Theta}$,

$$|\Delta_{J(\theta^*)^c}|_1 \leqslant c_0 |\Delta_{J(\theta^*)}|_1, \tag{2.9}$$

where $\Delta = \hat{\theta} - \theta^*$, $c_0 = 1$ *for the Dantzig estimator and* $c_0 = 3$ *for the Lasso.*

Proof. Define the random variables $Z_j = n^{-1}\sum_{i=1}^n X_{i,j}W_i$, $1 \leqslant j \leqslant M$. Using (2.5) we get that $Z_j \sim \mathcal{N}(0, \sigma^2/n)$, $1 \leqslant j \leqslant M$. Define the event

$$\mathcal{A} = \bigcap_{j=1}^M \{|Z_j| \leqslant r/2\}.$$

Standard inequalities on the tail of Gaussian variables yield

$$\begin{aligned} P(\mathcal{A}^c) &\leqslant MP(|Z_1| \geqslant r/2), \\ &\leqslant M \exp\left(-\frac{n}{2\sigma^2}\left(\frac{r}{2}\right)^2\right) \\ &\leqslant M^{1-\frac{A^2}{8}}. \end{aligned}$$

On the event $\mathcal{A}$, we have

$$\left|\frac{1}{n}X^T W\right|_\infty \leqslant \frac{r}{2}. \tag{2.10}$$

Any vector $\hat{\theta}$ in $\hat{\Theta}^L$ or $\hat{\Theta}^D$ satisfies the Dantzig constraint (2.4). Thus we have on $\mathcal{A}$ that

$$\sup_{\hat{\theta}\in\hat{\Theta}} \left|\Psi(\theta^* - \hat{\theta})\right|_\infty \leqslant \frac{3r}{2}.$$

Now we prove the second inequality. For any $\hat{\theta}^D \in \hat{\Theta}^D$, we have by definition that $|\hat{\theta}^D|_1 \leqslant |\theta^*|_1$, thus

$$|\Delta_{J(\theta^*)^c}|_1 = \sum_{j\in(J(\theta^*))^c} |\hat{\theta}^D_j| \leqslant \sum_{j\in J(\theta^*)} |\theta^*_j| - |\hat{\theta}^D_j| \leqslant |\Delta_{J(\theta^*)}|_1.$$

Consider now the Lasso estimators. By definition, we have for any $\hat{\theta}^L \in \hat{\Theta}^L$

$$\frac{1}{n}|Y - X\hat{\theta}^L|_2^2 + 2r|\hat{\theta}^L|_1 \leqslant \frac{1}{n}|W|_2^2 + 2r|\theta^*|_1.$$

Developing the left hand side on the above inequality, we get

$$2r|\hat{\theta}^L|_1 \leqslant 2r|\theta^*|_1 + \frac{2}{n}(\hat{\theta}^L - \theta^*)^T X^T W.$$

On the event $\mathcal{A}$, we have for any $\hat{\theta}^L \in \hat{\Theta}^L$

$$2|\hat{\theta}^L|_1 \leqslant 2|\theta^*|_1 + |\hat{\theta}^L - \theta^*|_1,$$

Adding $|\hat{\theta}^L - \theta^*|_1$ on both side, we get

$$\begin{aligned} |\hat{\theta}^L - \theta^*|_1 + 2|\hat{\theta}^L|_1 &\leqslant 2|\theta^*|_1 + 2|\hat{\theta}^L - \theta^*|_1 \\ |\hat{\theta}^L - \theta^*|_1 &\leqslant 2(|\hat{\theta}^L - \theta^*|_1 + |\theta^*|_1 - |\hat{\theta}^L|_1), \end{aligned}$$

Now we remark that if $j \in J(\theta^*)^c$, then we have $|\hat{\theta}^L_j - \theta^*_j| + |\theta^*_j| - |\hat{\theta}^L_j| = 0$. Thus we have on the event $\mathcal{A}$ that

$$\begin{aligned} |\Delta_{J(\theta^*)^c}|_1 - |\Delta_{J(\theta^*)}|_1 &\leqslant |\Delta|_1 \leqslant 2|\Delta_{J(\theta^*)}|_1 \\ |\Delta_{J(\theta^*)^c}|_1 &\leqslant 3|\Delta_{J(\theta^*)}|_1, \end{aligned}$$

for any $\hat{\theta}^L \in \hat{\Theta}^L$.

□

Lemma 2.2. *Let Assumption 2.2 be satisfied. Then*

$$\kappa(s,c_0) \stackrel{\triangle}{=} \min_{J\subset\{1,\cdots,M\},|J|\leqslant s}\ \min_{\lambda\neq 0:|\lambda_{J^c}|_1\leqslant c_0|\lambda_J|_1} \frac{|X\lambda|_2}{\sqrt{n}|\lambda_J|_2} \geqslant \sqrt{1-\frac{1}{\alpha}} > 0.$$

Proof. For any subset J of $\{1,\ldots,M\}$ such that $|J| \leqslant s$ and $\lambda \in \mathbb{R}^M$ such that $|\lambda_{J^c}|_1 \leqslant c_0|\lambda_J|_1$, we have

$$\begin{aligned}
\frac{|X\lambda_J|_2^2}{n|\lambda_J|_2^2} &= 1+\frac{\lambda_J^T(\Psi-I_M)\lambda_J}{|\lambda_J|_2^2} \\
&\geqslant 1-\frac{1}{\alpha(1+2c_0)s}\sum_{i,j=1}^{M}\frac{|\lambda_J^{(i)}||\lambda_J^{(j)}|}{|\lambda_J|_2^2} \\
&\geqslant 1-\frac{1}{\alpha(1+2c_0)s}\frac{|\lambda_J|_1^2}{|\lambda_J|_2^2}, \qquad (2.11)
\end{aligned}$$

where we have used Assumption 2.2 in the second line, I_M denotes the $M\times M$ identity matrix and $\lambda_J = (\lambda_J^{(1)},\ldots,\lambda_J^{(M)})$ denotes the components of the vector λ_J. This yields

$$\begin{aligned}
\frac{|X\lambda|_2^2}{n|\lambda_J|_2^2} &\geqslant \frac{|X\lambda_J|_2^2}{n|\lambda_J|_2^2}+2\frac{\lambda_J^TX^TX\lambda_{J^c}}{n|\lambda_J|_2^2} \\
&\geqslant 1-\frac{1}{\alpha s(1+2c_0)}\frac{|\lambda_J|_1^2}{|\lambda_J|_2^2}-\frac{2}{\alpha s(1+2c_0)}\frac{|\lambda_J|_1|\lambda_{J^c}|_1}{|\lambda_J|_2^2} \\
&\geqslant 1-\frac{1}{\alpha s}\frac{|\lambda_J|_1^2}{|\lambda_J|_2^2} \\
&\geqslant 1-\frac{1}{\alpha} > 0.
\end{aligned}$$

We have used Assumption 2.2 in the second line, the inequality $|\lambda_{J^c}|_1 \leqslant c_0|\lambda_J|_1$ in the third line and the fact that $|\lambda_J|_1 \leqslant \sqrt{|J|}|\lambda_J|_2 \leqslant \sqrt{s}|\lambda_J|_2$ in the last line.

□

Proof of Theorem 2.1. For all $1\leqslant j\leqslant M$, $\hat\theta\in\hat\Theta$ we have

$$(\Psi(\theta^*-\hat\theta))_j = (\theta_j^*-\hat\theta_j)+\sum_{i=1,i\neq j}^{M}\Psi_{i,j}(\theta_i^*-\hat\theta_i).$$

Assumption 2.2 yields

$$|(\Psi(\theta^*-\hat\theta))_j-(\theta_j^*-\hat\theta_j)| \leqslant \frac{1}{\alpha(1+2c_0)s}\sum_{i=1,i\neq j}^{M}|\theta_i^*-\hat\theta_i|,\ \forall j.$$

Thus we have

$$|\theta^* - \hat{\theta}|_\infty \leqslant \left|\Psi(\theta^* - \hat{\theta})\right|_\infty + \frac{1}{\alpha(1+2c_0)s}|\theta^* - \hat{\theta}|_1. \tag{2.12}$$

Set $\Delta = \hat{\theta} - \theta^*$. Lemma 2.1 yields that on an event $\mathcal{A}$ of probability at least $1 - M^{1-A^2/8}$ we have for any $\hat{\theta} \in \hat{\Theta}$

$$|\Psi\Delta|_\infty \leqslant \frac{3r}{2}, \tag{2.13}$$

and

$$|\Delta|_1 = |\Delta_{J(\theta^*)^c}|_1 + |\Delta_{J(\theta^*)}|_1 \leqslant (1+c_0)|\Delta_{J(\theta^*)}|_1 \leqslant (1+c_0)\sqrt{s}|\Delta_{J(\theta^*)}|_2.$$

Thus we have, on the same event $\mathcal{A}$,

$$\begin{aligned} \frac{1}{n}|X\Delta|_2^2 &= \Delta^T\Psi\Delta \\ &\leqslant |\Psi\Delta|_\infty|\Delta|_1 \\ &\leqslant \frac{3r}{2}(1+c_0)\sqrt{s}|\Delta_{J(\theta^*)}|_2, \end{aligned} \tag{2.14}$$

for any $\hat{\theta} \in \hat{\Theta}$. Lemma 2.2 yields

$$\frac{1}{n}|X\Delta|_2^2 \geqslant \left(1 - \frac{1}{\alpha}\right)|\Delta_{J(\theta^*)}|_2^2, \tag{2.15}$$

for any $\hat{\theta} \in \hat{\Theta}$. Combining (2.14) and (2.15), we obtain that

$$|\Delta|_1 \leqslant \frac{3}{2}r(1+c_0)^2\frac{\alpha}{\alpha-1}s, \tag{2.16}$$

for any $\hat{\theta} \in \hat{\Theta}$. Combining (2.12), (2.13) and (2.16) we obtain that

$$\sup_{\hat{\theta}\in\hat{\Theta}}|\hat{\theta} - \theta^*|_\infty \leqslant \frac{3}{2}\left(1 + \frac{(1+c_0)^2}{(1+2c_0)(\alpha-1)}\right)r. \square$$

Proof of Theorem 2.2. Theorem 2.1 yields $\sup_{\hat{\theta}\in\hat{\Theta}}|\hat{\theta} - \theta^*|_\infty \leqslant c_2 r$ on an event $\mathcal{A}$ of probability at least $1 - M^{1-A^2/8}$. Take $\hat{\theta} \in \hat{\Theta}$. For $j \in J(\theta^*)^c$, we have $\theta_j^* = 0$, and $|\hat{\theta}_j| \leqslant c_2 r$ on $\mathcal{A}$. For $j \in J(\theta^*)$, we have by Assumption 2.3 that $|\theta_j^*| \geqslant c_1 r$ and $|\theta_j^*| - |\hat{\theta}_j| \leqslant |\theta_j^* - \hat{\theta}_j| \leqslant c_2 r$ on $\mathcal{A}$. Since we assume that $c_1 > 2c_2$, we have on $\mathcal{A}$ that $|\hat{\theta}_j| \geqslant (c_1 - c_2)r > c_2 r$. Thus on the event $\mathcal{A}$ we have: $j \in J(\theta^*) \Leftrightarrow |\hat{\theta}_j| > c_2 r$. This yields $\mathrm{sign}(\tilde{\theta}_j) = \mathrm{sign}(\hat{\theta}_j) = \mathrm{sign}(\theta_j^*)$ if $j \in J(\theta^*)$ on the event $\mathcal{A}$. If $j \notin J(\theta^*)$, $\mathrm{sign}(\theta_j^*) = 0$ and $\tilde{\theta}_j = 0$ on $\mathcal{A}$, so that $\mathrm{sign}(\tilde{\theta}_j) = 0$. The same reasoning holds true simultaneously for all $\hat{\theta} \in \hat{\Theta}$ on the event $\mathcal{A}$. Thus we get the result. $\square$

Proof of Theorem 2.3. The proof of Theorem 2.3 is similar to the one of Theorem 2.1 up to a modification of the bound on $P(\mathcal{A}^c)$ in Lemma 2.1. Recall that $Z_j = n^{-1}\sum_{i=1}^n X_{i,j}W_i$, $1 \leqslant j \leqslant M$ and the event $\mathcal{A}$ is defined by

$$\mathcal{A} = \bigcap_{j=1}^{M}\{|Z_j| \leqslant r/2\} = \{\max_{1\leqslant j\leqslant M}|Z_j| \leqslant r/2\}.$$

The Markov inequality yields that

$$P(\mathcal{A}^c) \leqslant \frac{4\mathbb{E}[\max_{1\leqslant j\leqslant M} Z_j^2]}{r^2}.$$

Then we use Lemma 2.3 given below with $p = \infty$ and the random vectors

$$Y_i = (X_{i,1}W_i/n, \ldots, X_{i,M}W_i/n) \in \mathbb{R}^M,\ i = 1, \ldots, n.$$

We get that

$$P(\mathcal{A}^c) \leqslant \tilde{c}\frac{\log M}{r^2}\sigma^2\sum_{i=1}^n \max_{1\leqslant j\leqslant M}\frac{X_{i,j}^2}{n^2},$$

where $\tilde{c} > 0$ is an absolute constant. Taking $r = \sigma\sqrt{(\log M)^{1+\delta}/n}$ and using Assumption 2.5 yields that

$$P(\mathcal{A}^c) \leqslant \frac{c}{(\log M)^{\delta}},$$

where $c > 0$ is an absolute constant. □

The following result is Lemma 5.2.2, page 188 of [84].

Lemma 2.3. *Let $Y_1, \ldots, Y_n \in \mathbb{R}^M$ be independent random vectors with zero means and finite variance, and let $M \geqslant 3$. Then for every $p \in [2, \infty]$, we have*

$$\mathbb{E}\left[|\sum_{i=1}^n Y_i|_p^2\right] \leqslant \tilde{c}\min[p, \log M]\sum_{i=1}^n \mathbb{E}\left[|Y_i|_p^2\right],$$

where $\tilde{c} > 0$ is an absolute constant.

Chapter 3

Assumptions on the design matrix for the estimation problem

In this chapter, we review some commonly used assumptions on the design matrix X in linear regression to derive estimation and variable selection consistency of the Lasso and the Dantzig Selector. We establish the connections between these assumptions. Juditsky and Nemirovski [56] establish the necessary and sufficient condition for exact reconstruction of the target vector in the noiseless case by constrained l_1 minimization. We provide in this chapter a direct proof of this result different from that in [56]. We also derive an l_1-estimation result for the Lasso and the Dantzig Selector in the presence of noise under this condition.

3.1 Introduction

Consider the linear regression model

$$Y = X\theta^* + W, \tag{3.1}$$

where X is an $n \times M$ deterministic design matrix, $\theta^* \in \mathbb{R}^M$ and $W = (W_1, \ldots, W_n)^T$ is a zero-mean random vector such that $\mathbb{E}[W_i^2] \leqslant \sigma^2$, $1 \leqslant i \leqslant n$ for some $\sigma^2 > 0$. The Gram matrix of the design is defined as follows

$$\Psi = \frac{1}{n} X^T X.$$

For any $\theta \in \mathbb{R}^M$, define $J(\theta) = \{j \ : \ \theta_j \neq 0\}$. Let $M(\theta) = |J(\theta)|$ be the cardinality of $J(\theta)$. We say a vector θ is s-sparse if $M(\theta) \leqslant s$. Define the sign vector of θ by $\overrightarrow{\text{sign}}(\theta) = (\text{sign}(\theta_1), \ldots, \text{sign}(\theta_M))^T$ where

$$\text{sign}(t) = \begin{cases} 1 & \text{if } t > 0, \\ 0 & \text{if } t = 0, \\ -1 & \text{if } t < 0. \end{cases}$$

For any vector $\theta \in \mathbb{R}^M$ and any subset J of $\{1, \ldots, M\}$, we denote by θ_J the vector in $\mathbb{R}^M$ which has the same coordinates as θ on J and zero coordinates on the complement J^c of J and by $\overline{\theta}_J$ the vector in $\mathbb{R}^J$ obtained by keeping only the components θ_j with index $j \in J$. For any subset J of $\{1, \ldots, M\}$, we denote by X_J the $n \times |J|$ sub-matrix of X obtained by keeping only the columns of X with their index in J. For any $j \in \{1, \ldots, M\}$, we denote by X_j the j-th column of X. For any integers $1 \leqslant d, p < \infty$ and $z = (z_1, \ldots, z_d) \in \mathbb{R}^d$, the l_p norm of the vector z is denoted by $|z|_p \triangleq \left(\sum_{j=1}^d |z_j|^p\right)^{1/p}$, and $|z|_\infty \triangleq \max_{1 \leqslant j \leqslant d} |z_j|$.

In this chapter, we study the question of possibility to recover the vector θ^* by the Lasso and the Dantzig Selector. Note that the assumption of uniqueness of θ^* is not satisfied if $M > n$. In this case, if a vector $\theta^* = \theta^0$ satisfies (3.1), then there exists an affine space $\Theta^* = \{\theta^* \ : \ X\theta^* = X\theta^0\}$ of dimension larger than $M - n$ of vectors satisfying (3.1). However we will see that under some proper assumption on the Gram matrix of the design, any s-sparse vector θ^* satisfying (3.1) is identifiable and can be recovered by the Lasso and the Dantzig Selector.

Recall that the Lasso and the Dantzig Selector $\hat{\theta}^L$, $\hat{\theta}^D$ solve respectively the minimization problems

$$\min_{\theta \in \mathbb{R}^M} \frac{1}{n} |Y - X\theta|_2^2 + 2r|\theta|_1, \tag{3.2}$$

and

$$\min_{\theta \in \mathbb{R}^M} |\theta|_1 \text{ subject to } \left| \frac{1}{n} X^T (Y - X\theta) \right|_\infty \leqslant r, \tag{3.3}$$

where $r = A\sigma\sqrt{(\log M)/n}$, for some $A > 0$. We denote respectively by $\hat{\Theta}^L$ and $\hat{\Theta}^D$ the set of solutions to the Lasso and Dantzig minimization problems (3.2) and (3.3).

We state now Lemma 1 from Rosenbaum and Tsybakov [91].

Lemma 3.1. *Let $\hat{\theta}$ be a solution of the problem $\min_{\theta \in \Theta'} |\theta|_1$, where Θ' is a subset of $\mathbb{R}^M$. Let θ^* be any element of Θ' and J any subset of $\{1, \ldots, M\}$. Then for $\Delta = \hat{\theta} - \theta^*$ we have*

$$|\Delta_{J^c}|_1 \leqslant |\Delta_J|_1 + 2|\theta^*_{J^c}|_1. \tag{3.4}$$

Consider the Dantzig Selector. In this case, the set Θ' is given by the constraint in (3.3) and if we consider $J = J(\theta^*)$ we obtain that

$$|\Delta_{J(\theta^*)^c}|_1 \leqslant |\Delta_{J(\theta^*)}|_1.$$

This property of the Dantzig Selector is well known, see e.g. [6, 16]. More generally, following [6], we can state a similar result for the Lasso estimator. We have, with probability close to 1, that for any solution $\hat{\theta}^L$ of the Lasso minimization problem

$$|\Delta_{J(\theta^*)^c}|_1 \leqslant c_0 |\Delta_{J(\theta^*)}|_1,$$

with typically $c_0 \geqslant 1$, where $\Delta = \hat{\theta}^L - \theta^*$ (see also Lemma 1 in Chapter 1 for a proof with $c_0 = 3$). The two above inequalities are key arguments in the proof of the exact reconstruction property of the Lasso and the Dantzig Selector as we will see it in Section 3.3.

Accurate estimation of the vector of unknown parameters via l_1 minimization requires conditions on the design matrix. Numerous sufficient conditions were proposed in the literature. In Section 3.2, we present some of them and overview the state of the art. In Section 3.3, we state the necessary and sufficient condition for exact reconstruction of any s-sparse vector θ^* in the noiseless case, see [30, 56, 35]. We also provide a direct proof of this result different from that in [56]. In Section 3.4, we treat the presence of noise. In this case, the exact reconstruction of the s-sparse vectors is no longer guaranteed but we can prove the estimation consistency of the Lasso and Dantzig estimators to any s-sparse vector θ^* under the same necessary and sufficient assumption on the design matrix as in the noiseless case.

3.2 Some sufficient assumptions on the design matrix

For any $j \in \{1, \ldots, M\}$, denote by X_j the j-th column of X. For any $J \subset \{1, \ldots, M\}$, denote by X_J the $n \times |J|$ sub-matrix of X obtained by keeping the columns X_j of X with index $j \in J$. For $1 \leqslant u \leqslant M$ define the following "restricted" eigenvalues:

$$\phi_{\min}(u) = \min_{x \in \mathbb{R}^M : 1 \leqslant M(x) \leqslant u} \frac{x^T \Psi x}{|x|_2^2}.$$

We will call the first condition we consider the Sparse Positive Definiteness Condition. Actually this condition is not sufficient to prove that any s-sparse vector can be recovered by l_1 minimization procedures such as the Lasso or the Dantzig Selector.

Assumption 3.1. *Let $s \geqslant 1$ be an integer. The $n \times M$ matrix X satisfies the Sparse Positive Definiteness Condition $\boldsymbol{SPD}(s)$ if there exists a constant $c > 0$ such that*

$$\phi_{\min}(s) \geqslant c > 0.$$

We prove in Lemma 3.2 below that condition $\mathbf{SPD}(2s)$ is necessary to ensure recovery of any s-sparse vector satisfying (3.1) with the Lasso and the Dantzig Selector. Note that condition $\mathbf{SPD}(2s)$ guarantees the identifiability of any s-sparse vector θ^* satisfying (3.1).

The second important condition is the Irrepresentable Condition.

Assumption 3.2. *Let $s \geqslant 1$ be an integer. The $n \times M$ matrix X satisfies the Irrepresentable Condition $\boldsymbol{IC}(s, \eta)$ if for any subset I of $\{1, \ldots, M\}$ with $|I| \leqslant s$*

- *the matrix X_I has trivial kernel,*
- *for any vector z in $\{-1, 1\}^{|I|}$, we have*

$$|X_{I^c}^T X_I (X_I^T X_I)^{-1} z|_\infty < \eta. \tag{3.5}$$

The Strong Irrepresentable Condition, i.e., the condition $\mathbf{IC}(s, \eta)$ with $0 \leqslant \eta < 1$ was used in many papers, see [3, 82, 116, 104], to derive asymptotic variable selection consistency of the Lasso. This condition appeared earlier in the articles on compressed sensing [35, 42]. In Chapter 4, we derive non-asymptotic rates of l_∞ estimation and sign concentration under the Strong Irrepresentable Condition $\mathbf{IC}(s, 1 - \gamma)$, for some $\gamma > 0$.

The next condition is called the Restricted Eigenvalue Condition $\mathbf{RE}(s, c_0)$. It was introduced in [6].

Assumption 3.3. *Let $s \geqslant 1$ be an integer and $c_0 > 0$. The $n \times M$ matrix X satisfies condition* ***RE***(s, c_0) *if*

$$\kappa(s, c_0) \stackrel{\triangle}{=} \min_{J_0 \subset \{1,\ldots,M\}: |J_0| \leqslant s} \quad \min_{\Delta \neq 0: |\Delta_{J_0^c}|_1 \leqslant c_0 |\Delta_{J_0}|_1} \frac{|X\Delta|_2}{\sqrt{n}|\Delta_{J_0}|_2} > 0.$$

For $1 \leqslant m, m' \leqslant M$ introduce the "restricted" correlations

$$\gamma_{m,m'} = \max_{J \cap J' = \emptyset, |J| \leqslant m, |J'| \leqslant m'} \left\{ \frac{1}{n} \theta^T X_J^T X_{J'} \theta' \, : \, \theta \in \mathbb{R}^m, \, \theta' \in \mathbb{R}^{m'}, \, |\theta|_2 \leqslant 1, \, |\theta'|_2 \leqslant 1 \right\}.$$

Assumption 3.4. *Let $s \geqslant 1$ be an integer and $c_0 > 0$. The $n \times M$ matrix X satisfies condition* ***RI***(s, c_0) *if for any $\theta \in \mathbb{R}^M$ with at most s nonzero components, we have*

$$\phi_{\min}(2s) > c_0 \gamma_{s,2s}$$

for some integer $1 \leqslant s \leqslant M/2$.

The condition $\mathbf{RI}(s, c_0)$ reduces to the Restricted Isometry condition of [16] when $c_0 = 1$. This condition is similar to the $\mathbf{RE}(s, c_0)$ condition but is more restrictive.

The last condition we present is the mutual coherence condition $\mathbf{MC}(s, c_0)$. It was first introduced in [34]. We state here the version of [71]. Denote by $\Psi = \frac{1}{n} X^T X$ the Gram matrix of the design.

Assumption 3.5. *Let $s \geqslant 1$ be an integer and $c_0 > 0$. We say that the mutual coherence condition* ***MC***(s, c_0) *is satisfied if the elements $\Psi_{i,j}$ of the Gram matrix Ψ satisfy*

$$\Psi_{j,j} = 1, \ \ \forall 1 \leqslant j \leqslant M, \tag{3.6}$$

and

$$\max_{i \neq j} |\Psi_{i,j}| \leqslant \frac{1}{\alpha(1 + 2c_0)s}, \tag{3.7}$$

for some integer $s \geqslant 1$ and some constant $\alpha > 1$.

For any integer $d \geqslant 1$ and any $d \times d$ matrix A, the standard minimum and maximum eigenvalues of A are defined respectively by

$$\Phi_{\min}(A) = \min_{x \in \mathbb{R}^d} \frac{x^T A x}{|x|_2^2},$$

and

$$\Phi_{\max}(A) = \max_{x \in \mathbb{R}^d} \frac{x^T A x}{|x|_2^2}.$$

For any subset $J \subset \{1, \ldots, M\}$ define $\Psi_{J,J} = \frac{X_J^T X_J}{n}$. We now give some implications between these conditions.

Theorem 3.1. *For any integer s such that $1 \leqslant s \leqslant n$, we have*

$$\boldsymbol{MC}(s, c_0) \Rightarrow \boldsymbol{RI}(s, c_0) \Rightarrow \boldsymbol{RE}(s, c_0), \tag{3.8}$$

$$\boldsymbol{RE}(s, c_0) \Rightarrow \boldsymbol{SPD}(2s), \quad \text{if } c_0 \geqslant 1, \tag{3.9}$$

$$\boldsymbol{MC}(s, c_0) \Rightarrow \boldsymbol{IC}\left(s, \frac{1}{\alpha(1+2c_0)-1}\right), \tag{3.10}$$

where $\alpha > 0$ is defined in $\boldsymbol{MC}(s, c_0)$, and

$$\boldsymbol{IC}(s, \eta) \Rightarrow \boldsymbol{SPD}(2s), \quad \forall 0 \leqslant \eta \leqslant 1. \tag{3.11}$$

In addition, if

$$\eta < \frac{1}{2c_0(1+\alpha)s} \min_{J \subset \{1,\ldots,M\}: |J| \leqslant s} \left(\frac{\Phi_{\min}(\Psi_{J,J})}{\Phi_{\max}(\Psi_{J,J})} \right), \tag{3.12}$$

for some $\alpha > 0$, then

$$\boldsymbol{IC}(s, \eta) \Rightarrow \boldsymbol{RE}(s, c_0).$$

Proof. We prove that $\mathbf{MC}(s, c_0) \Rightarrow \mathbf{RI}(s, c_0)$.

For any $\theta \neq 0$ with at most s nonzero components, we have

$$\begin{aligned} \frac{|X\theta|_2^2}{n|\theta|_2^2} &= \frac{\theta^T \Psi \theta}{|\theta|_2^2} \\ &= 1 + \frac{\theta^T(\Psi - I_M)\theta}{|\theta|_2^2}. \end{aligned}$$

Thus

$$\frac{|X\theta|_2^2}{n|\theta|_2^2} \geqslant 1 - \frac{1}{cs} \frac{|\theta|_1^2}{|\theta|_2^2},$$

where $c = \alpha(1 + 2c_0)$. The Cauchy-Schwarz inequality yields

$$\frac{|X\theta|_2^2}{n|\theta|_2^2} \geqslant 1 - \frac{1}{c}.$$

Next, by definition of $\gamma_{s,2s}$ we have for any subsets of indices J and J' such that $J \cap J' = \emptyset$, $|J| \leqslant s$ and $|J'| \leqslant 2s$

$$\begin{aligned} c_0 \gamma_{s,2s} &\leqslant \frac{c_0}{cs} \max\{|\theta_J|_1 |\theta_{J'}|_1 \,:\, \theta \in \mathbb{R}^M,\, |\theta_J|_2 \leqslant 1,\, |\theta_{J'}|_2 \leqslant 1,\, |J| \leqslant s,\, |J'| \leqslant 2s\} \\ &\leqslant \frac{\sqrt{2}c_0}{c} \\ &\leqslant 1 - \frac{1}{c}, \end{aligned}$$

by definition of c and Assumption 3.5. This yields the first result.

Next, it is proved in [6] that $\mathbf{RI}\,(s, c_0) \Rightarrow \mathbf{RE}(s, c_0)$, see Lemma 3 in [6]. The implication (3.9) is also proved in [6] page 1710.

We prove now the implication (3.10). Assume that $\mathbf{MC}\,(s, c_0)$ is satisfied. Then, combining (3.8) and (3.9) yields the first point of $\mathbf{IC}(s, \eta)$. Next, for any subset J in $\{1, \ldots, M\}$ such that $|J| \leqslant s$ and any vector $z \in \{-1, +1\}^{|J|}$ we have

$$|\Psi_{J^c,J}\Psi_{J,J}^{-1}z|_\infty \leqslant \frac{1}{\alpha(1+2c_0)s}|\Psi_{J,J}^{-1}z|_1. \tag{3.13}$$

Define the matrices $U = \Psi - I_M$ where I_M is the $M \times M$ identity matrix. Denote the components of the matrix U by $U_{i,j}$, $1 \leqslant i, j \leqslant M$. We have that $\max_{i,j}|U_{i,j}| \leqslant \frac{1}{\alpha(1+2c_0)s}$. We have $\Psi_{J,J} = I_{|J|} + U_{J,J}$ where $I_{|J|}$ denotes the $|J| \times |J|$ identity matrix. We have

$$\Psi_{J,J}^{-1} = (I_{|J|} + U_{J,J})^{-1} = I_{|J|} + \tilde{U},$$

where $\tilde{U} = \sum_{k=1}^{\infty}(-1)^k U_{J,J}^k$. Note that this series converges since for any $k \in \mathbb{N}$, the elements of $U_{J,J}^k$ are bounded in absolute value by $\frac{1}{(\alpha(1+2c_0))^k s} < 1$ if the condition $\mathbf{MC}\,(s, c_0)$ holds. Assumption $\mathbf{MC}\,(s, c_0)$ yields that

$$\max_{i,j}|\tilde{U}_{i,j}| \leqslant \frac{1}{s(\alpha(1+2c_0)-1)}. \tag{3.14}$$

Combining (3.13), (3.14) yields

$$|\Psi_{J^c,J}\Psi_{J,J}^{-1}z|_\infty \leqslant \frac{1}{\alpha(1+2c_0)-1}.$$

The implication (3.11) is obtained by combining Lemma 3.2 and Theorem 3.3 below.

We now prove the last implication. For any subset J of $\{1, \ldots, M\}$ such that $|J| \leqslant s$ and $\Delta \in \mathbb{R}^M \setminus \{0\}$ such that $|\Delta_{J^c}|_1 \leqslant c_0|\Delta_J|_1$ we have

$$\begin{aligned}
\frac{\Delta^T\Psi\Delta}{|\Delta_J|_2} &= \frac{\overline{\Delta}_J^T\Psi_{J,J}\overline{\Delta}_J}{|\Delta_J|_2} + 2\frac{\overline{\Delta}_{J^c}^T\Psi_{J^c,J}\overline{\Delta}_J}{|\Delta_J|_2} + \frac{\overline{\Delta}_{J^c}^T\Psi_{J^c,J^c}\overline{\Delta}_{J^c}}{|\Delta_J|_2} \\
&\geqslant \Phi_{\min}(\Psi_{J,J}) - 2\max_{\overline{\Delta}_J\in\mathbb{R}^{|J|}}\left(\frac{|\overline{\Delta}_{J^c}^T\Psi_{J^c,J}\overline{\Delta}_J|}{|\overline{\Delta}_J|_2^2}\right) \\
&\geqslant \Phi_{\min}(\Psi_{J,J}) - 2\max_{\overline{\Delta}_J\in\mathbb{R}^{|J|}}\left(\frac{|\overline{\Delta}_{J^c}|_1|\Psi_{J^c,J}\overline{\Delta}_J|_\infty}{|\overline{\Delta}_J|_2^2}\right) \\
&\geqslant \Phi_{\min}(\Psi_{J,J}) - 2c_0\sqrt{s}\max_{\overline{\Delta}_J\in\mathbb{R}^{|J|}}\left(\frac{|\Psi_{J^c,J}\Psi_{J,J}^{-1}\Psi_{J,J}\overline{\Delta}_J|_\infty}{|\overline{\Delta}_J|_2}\right).
\end{aligned} \tag{3.15}$$

Denote by u the vector of $\mathbb{R}^M$ such that $u_{J^c} = 0$ and

$$u_j = \frac{\Psi_{j,J}\overline{\Delta}_J}{|\Psi_{J,J}\overline{\Delta}_J|_1},\ \forall j \in J.$$

We have

$$\begin{aligned}
\max_{\overline{\Delta}_J \in \mathbb{R}^{|J|}} \left(\frac{|\Psi_{J^c,J} \Psi_{J,J}^{-1} \Psi_{J,J} \overline{\Delta}_J|_\infty}{|\overline{\Delta}_J|_2} \right) &= \max_{\overline{\Delta}_J \in \mathbb{R}^{|J|}} \left(\frac{|\Psi_{J^c,J} \Psi_{J,J}^{-1} \overline{u}_J|_\infty |\Psi_{J,J} \overline{\Delta}_J|_1}{|\overline{\Delta}_J|_2} \right) \\
&\leqslant \max_{\overline{u}_J \in \mathbb{R}^{|J|} \,:\, |\overline{u}_J|_1 \leqslant 1,\, |\overline{u}_J|_\infty \leqslant 1} \left(|\Psi_{J^c,J} \Psi_{J,J}^{-1} \overline{u}_J|_\infty \right) \max_{\Delta_J \in \mathbb{R}^{|J|}} \left(\frac{|\Psi_{J,J} \overline{\Delta}_J|_1}{|\overline{\Delta}_J|_2} \right) \\
&\leqslant \sqrt{s} \max_{\overline{u}_J \in \mathbb{R}^{|J|} \,:\, |\overline{u}_J|_\infty \leqslant 1} \left(|\Psi_{J^c,J} \Psi_{J,J}^{-1} \overline{u}_J|_\infty \right) \max_{\overline{\Delta}_J \in \mathbb{R}^{|J|}} \left(\frac{|\Psi_{J,J} \overline{\Delta}_J|_2}{|\overline{\Delta}_J|_2} \right) \\
&\leqslant \sqrt{s} \max_{z \in \{-1,1\}^{|J|}} \left(|\Psi_{J^c,J} \Psi_{J,J}^{-1} z|_\infty \right) \Phi_{\max}(\Psi_{J,J}) \\
&\leqslant \sqrt{s} \eta \Phi_{\max}(\Psi_{J,J}), \qquad (3.16)
\end{aligned}$$

where we have used that $\max_{\overline{u}_J \in \mathbb{R}^{|J|} \,:\, |\overline{u}_J|_\infty \leqslant 1} \left(|\Psi_{J^c,J} \Psi_{J,J}^{-1} \overline{u}_J|_\infty \right) = \max_{z \in \{-1,1\}^{|J|}} \left(|\Psi_{J^c,J} \Psi_{J,J}^{-1} z|_\infty \right)$ and the inequality (3.12) in the last line. Combining (3.15) and (3.16) yields

$$\begin{aligned}
\frac{\Delta^T \Psi \Delta}{|\Delta_J|_2} &\geqslant \Phi_{\min}(\Psi_{J,J}) - 2c_0 s \eta \Phi_{\max}(\Psi_{J,J}) \\
&> \Phi_{\min}(\Psi_{J,J}) \frac{\alpha}{1+\alpha} > 0.
\end{aligned}$$

□

Unfortunately, we cannot fully describe the relations between the conditions $\mathbf{RE}(s, c_0)$ and $\mathbf{IC}(s, 1)$. Zou [117] gives a counter-example showing that condition $\mathbf{IC}(s, \eta)$ may not be satisfied for a positive definite Ψ. On the other hand, positive definiteness of Ψ implies $\mathbf{RE}(s, c_0)$. We believe that conditions $\mathbf{RE}(s, c_0)$ and $\mathbf{IC}(s, 1)$ are not related in a simple way.

3.3 Necessary and sufficient condition for exact reconstruction in the noiseless case

In the noiseless case, i.e., when $Y = X\theta^*$, the Lasso and Dantzig minimization problems are replaced by the following problem:

$$\min_{\theta \in \mathbb{R}^M} |\theta|_1 \text{ subject to } Y = X\theta.$$

It will be more convenient to write this minimization problem in the form:

$$\min_{\theta \in \mathbb{R}^M} |\theta|_1 \text{ subject to } X\theta^* = X\theta. \qquad (3.17)$$

Definition 3.1. *We say that the matrix X is s-good if, for any vector θ^* in $\mathbb{R}^M$ with at most s non-zero components, the unique solution of the problem (3.17) is θ^*.*

Finding the sparsest vector $\theta^* \in \mathbb{R}^M$ satisfying $Y = X\theta^*$ is of interest in practical applications. This motivates the study of the following problem in [21]:

$$\min_{\theta \in \mathbb{R}^M} M(\theta) \text{ subject to } Y = X\theta. \tag{3.18}$$

This is a non-convex combinatorial optimization problem because of the presence of the term $M(\theta)$. Hence, it is not computationally tractable when the dimension is high. Clearly the solution θ^* of (3.18) is also solution of (3.17) if X is s-good and if the vector θ^* has at most s nonzero components. This property is interesting because (3.17) is a convex optimization problem computationally tractable even when the dimension is high. Therefore, it is important to obtain characterizations of the property that X is s-good.

We introduce now some definitions. Fix $1 \leqslant s \leqslant M$. Define the polytope

$$P_s = \left\{ u \in \mathbb{R}^M \, : \, |u|_1 \leqslant s, \, |u|_\infty \leqslant 1 \right\},$$

and the quantity

$$\alpha_s(X) = \max_{u, \theta \in \mathbb{R}^M} \left\{ u^T\theta \, : \, u \in P_s, \, |\theta|_1 \leqslant 1, \, X\theta = 0 \right\}.$$

An interpretation of this quantity is in order. For any $\theta \in \mathbb{R}^M$ and any integer $1 \leqslant s \leqslant M$ define

$$|\theta|_{\max(s)} = \max_{J \subset \{1,\ldots,M\} \, : \, |J| \leqslant s} \sum_{j \in J} |\theta_j|.$$

The quantity $|\theta|_{\max(s)}$ is the sum of the s largest absolute values of the components of θ. The quantity $\alpha_s(X)$ is the maximum of the sums of the s largest absolute values of the components of the l_1-normalized vectors $\theta \in \ker(X)$. We have

$$\alpha_s(X) = \max_{\theta \, : \, |\theta|_1 \leqslant 1, \, X\theta = 0} \{ |\theta|_{\max(s)} \}.$$

We say that a s-sparse vector θ^* satisfying (3.1) is identifiable if there is no s-sparse vector θ' different from θ^* such that $X\theta' = X\theta^*$. This property is satisfied if condition ***SPD***(2s) is satisfied. The next lemma ensures that any s-sparse vector θ^* satisfying (3.1) is identifiable if the matrix X is s-good.

Lemma 3.2. *Assume that X is s-good with $1 \leqslant s \leqslant n$. Then Assumption* **SPD***(2s) is satisfied.*

Proof. Let I be a subset of $\{1,\dots,M\}$ such that $|I| = 2s$. Assume that X_I has a non-trivial kernel, i.e., there exists a nonzero vector θ in $\mathbb{R}^M$ such that $J(\theta) \subseteq I$ and $X\theta = 0$. We consider the case where $M(\theta) \geqslant s+1$ (if $M(\theta) \leqslant s$, then $\theta = 0$ since X is s-good). Consider arbitrary sets I_1 and I_2 in $\{1,\dots,M\}$ such that $|I_j| = s$ and $I_j \cap J(\theta) \neq \emptyset$ for $j = 1,2$, $I_1 \cap I_2 = \emptyset$ and $I = I_1 \cup I_2$. Define $\theta^{(j)} = \theta_{I_j}$ for $j = 1,2$. Note that $\theta = \theta^{(1)} + \theta^{(2)}$ and

$$X\theta^{(1)} = X(-\theta^{(2)}),$$

since $X\theta = 0$. Set $\theta^* = \theta^{(1)}$. Since $M(\theta^{(1)}) \leqslant s$ and X is s-good, the unique solution of the problem (3.17) is θ^*. This yields $|\theta^*|_1 < |\theta'|_1$ for any s-sparse vector $\theta' \in \mathbb{R}^M$ satisfying the constraint in (3.17). Note that the s-sparse vector $-\theta^{(2)}$ satisfies the constraint in (3.17). Thus we have $|\theta^{(1)}|_1 < |\theta^{(2)}|_1$. We set now $\theta^* = \theta^{(2)}$. The same reasoning as above yields $|\theta^{(2)}|_1 < |\theta^{(1)}|_1$. Thus X_I has trivial kernel.

□

Donoho and Tanner [31] and Juditsky and Nemirovski [56] derived the necessary and sufficient condition for θ^* to be the unique solution of the problem (3.17). This condition can be formulated as follows.

Theorem 3.2. *Fix $1 \leqslant s \leqslant M$. The matrix X is s-good if and only if $\alpha_s(X) < 1/2$.*

This condition says that for any l_1-normalized vector θ in $\ker(X)$, the sum of the s largest absolute values of the components of θ must be strictly smaller than $1/2$. Heuristically, this means that $\ker(X)$ cannot contain vectors whose l_1-norm is concentrated on a small set of components. Juditsky and Nemirovski [56] proved this result in their Theorem 2.2 and Corollary 2.1 by establishing that condition $\alpha_s(X) < 1/2$ is in fact a reformulation of the classical Karush-Kuhn-Tucker (KKT) condition for any s-sparse vector θ^* to be the unique solution of the minimization problem (3.17). For the sake of completeness, we provide below a simple and direct proof of Theorem 3.2 different from that in [56].

Proof. Assume that $\alpha_s(X) < 1/2$. Fix an arbitrary $\theta^* \in \mathbb{R}^M$ with at most s nonzero components. Let there exist a solution $\tilde{\theta} \in \mathbb{R}^M$ of the problem (3.17) distinct from θ^*. Set $\Delta = \tilde{\theta} - \theta^*$ and $J^* = J(\theta^*)$. By definition of $\tilde{\theta}$ and using Lemma 3.1 with $J = J^*$ we get

$$|\Delta_{J^{*c}}|_1 \;\leqslant\; |\Delta_{J^*}|_1.$$

By definition of $\tilde{\theta}$ we have $X\tilde{\theta} = X\theta^*$. Thus, $\Delta \in \ker(X)$. This combined with the definition of $\alpha_s(X)$ yields

$$|\Delta|_1 \leqslant 2|\Delta_{J^*}|_1 \leqslant 2|\Delta|_{\max(s)} \leqslant 2\alpha_s(X)|\Delta|_1 < |\Delta|_1,$$

since we assumed $\alpha_s(X) < 1/2$. This yields $\tilde{\theta} = \theta^*$. Thus, X is s-good.

Assume now that X is s-good and that $\alpha_s(X) \geqslant 1/2$. The set $P_s \times \{\theta : |\theta|_1 \leqslant 1, X\theta = 0\}$ is compact and the mapping $(u, \theta) \mapsto u^T\theta$ is continuous. Then, by definition of $\alpha_s(X)$ there exists a couple $(u_0, \theta_0) \in P_s \times \{\theta : |\theta|_1 \leqslant 1, X\theta = 0\}$ with $M(\theta_0) > s$ such that $u_0^T\theta_0 \geqslant 1/2$ (indeed, $M(\theta_0) \leqslant s$ is impossible because it implies $\theta_0 = 0$ since X is s-good). This yields $|\theta_0|_{\max(s)} \geqslant u_0^T\theta_0 \geqslant 1/2$. Denote by J the set of the s largest in absolute value components of θ_0. Set $\theta^* = (\theta_0)_J$. We have $|\theta^*|_1 = |\theta_0|_{\max(s)}$. Consider the vector $\theta' = \theta^* - \frac{1}{2}\theta_0$. Note that $\theta' \neq \theta^*$, $X\theta' = X\theta^*$ and $|\theta'|_1 = \frac{1}{2}|(\theta_0)_J - (\theta_0)_{J^c}|_1 \leqslant \frac{1}{2} \leqslant |\theta^*|_1$. This contradicts the fact that X is s-good. □

Now we prove that the Irrepresentable Condition is a sufficient condition for exact reconstruction with the procedure (3.17).

Theorem 3.3. *Fix $1 \leqslant s \leqslant M$. Let Assumption **IC**$(s, 1)$ be satisfied. Then the $n \times M$ matrix X is s-good.*

Proof. Let $\theta^* \in \mathbb{R}^M$ be a s-sparse vector. If $\theta^* = 0$, then it is trivially the unique solution of (3.17). Thus, it suffices to consider that $\theta^* \neq 0$. Set $I = J(\theta^*)$, $s' = |I| \leqslant s$ and $I = \{j_1, \ldots, j_{s'}\}$. Define $z = (\mathrm{sign}(\theta^*_{j_1}), \ldots, \mathrm{sign}(\theta^*_{j_{s'}}))^T$ and $u = X_I(X_I^T X_I)^{-1}z$. Since condition $\mathbf{IC}(s, 1)$ is satisfied we have

$$X_j^T u = \begin{cases} \mathrm{sign}(\theta^*_j), & \text{if } j \in I, \\ v_j, & \text{with } |v_j| < 1, \text{ if } j \in I^c, \end{cases} \tag{3.19}$$

where X_j is the jth column of X. Define the function $f : \mathbb{R}^M \to \mathbb{R}$ by

$$\begin{aligned} f(\theta) &= |\theta|_1 - u^T X(\theta - \theta^*) \\ &= \sum_{j \in I}[|\theta_j| - \mathrm{sign}(\theta^*_j)(\theta_j - \theta^*_j)] + \sum_{j \in I^c}[|\theta_j| - v_j\theta_j]. \end{aligned}$$

Now, assume that there exists a solution θ' of the problem (3.17) different from θ^*. Obviously, we have $f(\theta') = f(\theta^*)$, i.e.,

$$0 = f(\theta') - f(\theta^*) = \sum_{j \in I}[|\theta'_j| - \mathrm{sign}(\theta^*_j)(\theta'_j - \theta^*_j) - |\theta^*_j|] + \sum_{j \in I^c}[|\theta'_j| - v_j\theta'_j].$$

In the above display, the first sum on the right-hand-side is nonnegative by convexity of the function $x \to |x|$. Thus, we cannot have that $J(\theta') \cap I^c \neq \emptyset$, since then the second term is strictly positive in view of the condition $|v_j| < 1$ for any $j \in I^c$. Hence, we have $J(\theta') \subseteq I$. Then, we use that $X(\theta^* - \theta') = 0$ and X_I has trivial kernel to conclude that $\theta^* = \theta'$. □

Note that the Irrepresentable Condition $\mathbf{IC}(s,1)$ is more restrictive than the necessary and sufficient condition of exact recovery in the case $s < n$. Indeed, if $s < n$ then $\ker(X_I^T)$ is non trivial where $I = J(\theta^*)$. Juditsky and Nemirovski [56] proved that a necessary and sufficient condition for the s-sparse vector θ^* to be the unique solution of (3.17) is that there exists a vector $u \in \mathbb{R}^n$ satisfying (3.19). W.l.o.g. we can suppose that $X_I^T X_I$ is positive definite, see Lemma 3.2 and Theorem 3.3. Note that the vector u satisfying (3.19) can be expressed in the form

$$u = X_I(X_I^T X_I)^{-1} z + u',$$

with $z = (\mathrm{sign}(\theta^*_{j_1}), \ldots, \mathrm{sign}(\theta^*_{j_{|I|}}))^T$ and $u' \in \ker(X_I^T)$. In other words, the necessary and sufficient condition can be rewritten as:

$$\exists u' \in \ker(X_I^T) \,\mathrm{s.t.}\, |X_{I^c}^T X_I (X_I^T X_I)^{-1} z + X_{I^c}^T u'|_\infty < 1. \tag{3.20}$$

We see that (3.20) does not imply $\mathbf{IC}(s,1)$ since u' can be different from 0 if $s < n$. On the other hand, $\mathbf{IC}(s,1)$ implies (3.20). Note that Dossal [35] specifies additional assumptions on X and z such that condition $\mathbf{IC}(s,1)$ is the necessary and sufficient assumption for exact reconstruction with the procedure (3.17).

3.4 Estimation with Lasso and Dantzig Selector in the presence of noise

We now come back to the noisy model (3.1). In the presence of noise, exact reconstruction of the target vector θ^* is impossible. However, an interesting question is what kind of results can be obtained under the necessary and sufficient condition $\alpha_s(X) < 1/2$. The next theorem shows that for the Dantzig Selector an l_1 estimation result can be obtained under this condition. We also propose an l_1 estimation result for the Lasso under a more restrictive condition.

Theorem 3.4. *Take $r = A\sigma\sqrt{(\log M)/n}$ and $A > 2\sqrt{2}$. Fix $1 \leqslant s \leqslant n$. Define the quantity*

$$\overline{\beta} = \max_{z \in \{-1,0,1\}^M : M(z) \leqslant s} \min_{y \in \mathbb{R}^M} \{|y|_1 \,:\, |\Psi y - z|_\infty \leqslant \alpha_s(X)\}. \tag{3.21}$$

Assume that $\alpha_s(X) < 1/2$. If $M(\theta^*) \leqslant s$, then we have, with probability at least $1 - M^{1-\frac{A^2}{8}}$,

$$\sup_{\hat{\theta}^D \in \hat{\Theta}^D} |\hat{\theta}^D - \theta^*|_1 \leqslant \frac{3\overline{\beta}}{1 - 2\alpha_s(X)} r.$$

Assume now that $\alpha_s(X) < 1/4$. *If* $M(\theta^*) \leqslant s$, *then we have, with probability at least* $1 - M^{1-\frac{A^2}{8}}$,

$$\sup_{\hat{\theta}^L \in \hat{\Theta}^L} |\hat{\theta}^L - \theta^*|_1 \leqslant \frac{6\overline{\beta}}{1 - 4\alpha_s(X)} r.$$

Theorem 3.4 provides an oracle inequality for l_1 estimation of θ^* in the presence of noise under the weakest possible condition $\alpha_s(X) < 1/2$ for the Dantzig Selector and the more restrictive condition $\alpha_s(X) < 1/4$ for the Lasso. Unfortunately, the quantity $\overline{\beta}$ is not explicitly given. However, it is possible to compute numerical bounds on $\overline{\beta}$. Indeed, we have

$$\max_{z \in \{-1,0,1\}^M : M(z) \leqslant s} \min_{y \in \mathbb{R}^M} \{|y|_1 \, : \, |\Psi y - z|_\infty \leqslant \alpha_s(X)\} \leqslant \max_{u \in P_s} \min_{y \in \mathbb{R}^M} \{|y|_1 \, : \, |\Psi y - u|_\infty \leqslant \alpha_s(X)\}.$$

Using the computationally feasible bounds from above and below for the quantity $\alpha_s(X)$ of Juditsky and Nemirovski [56] and the fact that the constraints in the above display are convex, we can compute bounds from above on $\overline{\beta}$. Note that when the matrix Ψ is diagonal, we have the obvious bound $\overline{\beta} \leqslant s$. In the general case we believe that $\overline{\beta}$ depends also linearly on s.

The proof of Theorem 3.4 is inspired by [56] which considers the problem (3.17) in the presence of non-stochastic noise. The non-stochastic noise can be interpreted as bias and does not cover the case of stochastic noise considered here.

Proof. First, by Lemma 3.2 the condition $\alpha_s(X) < 1/2$ implies the condition $\mathbf{SPD}(2s)$ on X. Thus, $\Psi_{J,J}^{-1}$ is well defined for any $J \subset \{1, \ldots, M\}$ with $|J| \leqslant s$.

Next, for any $\beta > 0$ define the following quantities

$$\tilde{\alpha}_s(\Psi, \beta) = \max_{u, \theta \in \mathbb{R}^M} \{u^T \theta - \beta |\Psi \theta|_\infty, : u \in P_s, \, |\theta|_1 \leqslant 1\},$$

and

$$\tilde{\alpha}_s(\Psi) = \max_{u, \theta \in \mathbb{R}^M} \{u^T \theta, : u \in P_s, \, |\theta|_1 \leqslant 1, \, \Psi\theta = 0\}.$$

Note that since $\ker(X) = \ker(X^T X)$ we have $\tilde{\alpha}_s(\Psi) = \alpha_s(X^T X) = \alpha_s(X)$.

We have that $\tilde{\alpha_s}(X, \beta)$ is the infimum of $\alpha > 0$ such that for every vector $z \in \{-1, 0, 1\}^M$ with s nonzero components, there exists a vector $y \in \mathbb{R}^M$ such that

$$|y|_1 \leqslant \beta, \;\; \text{and} \;\; |\Psi y - z|_\infty \leqslant \alpha. \tag{3.22}$$

This follows from Theorem 2.2 in [56] applied to the matrix Ψ instead of the matrix X. Similarly, we have that $\tilde{\alpha}_s(\Psi)$ is the infimum of $\alpha > 0$ such that for every vector $z \in \{-1, 0, 1\}^M$ with s nonzero components, there exists a vector $y \in \mathbb{R}^M$ such that

$$|\Psi y - z|_\infty \leqslant \alpha. \tag{3.23}$$

Thus the quantity $\overline{\beta}$ is well defined and finite since the maximum is taken on a finite set in (3.21). By definition of $\tilde{\alpha}_s(\Psi, \beta)$ and $\tilde{\alpha}_s(\Psi)$ we have

$$\tilde{\alpha}_s(\Psi, \beta) \geqslant \tilde{\alpha}_s(\Psi), \ \forall \beta > 0.$$

Next, (3.22) and (3.23) imply that

$$\tilde{\alpha}_s(\Psi, \beta) = \tilde{\alpha}_s(\Psi), \ \forall \beta \geqslant \overline{\beta}.$$

Next, let the set $\hat{\Theta}$ denote either the set of Lasso solutions $\hat{\Theta}^L$ or Dantzig solutions $\hat{\Theta}^D$. For any $\hat{\theta} \in \hat{\Theta}$ set $\Delta = \hat{\theta} - \theta^*$ and $J^* = J(\theta^*)$.

It is proved in Lemma 2.1 of Section 2.4 that, with probability at least $1 - M^{1-\frac{A^2}{8}}$,

$$\sup_{\hat{\theta} \in \hat{\Theta}} \left| \Psi(\theta^* - \hat{\theta}) \right|_\infty \leqslant \frac{3r}{2}, \tag{3.24}$$

and for all $\hat{\theta} \in \hat{\Theta}$

$$|\Delta_{J^{*c}}|_1 \leqslant c_0 |\Delta_{J^*}|_1, \tag{3.25}$$

with $c_0 = 1$ for the Dantzig Selector and $c_0 = 3$ for the Lasso. Thus we have

$$\begin{aligned} |\Delta|_1 &= |\Delta_{J^{*c}}|_1 + |\Delta_{J^*}|_1 \\ &\leqslant (1 + c_0)|\Delta_{J^*}|_1 \\ &\leqslant (1 + c_0)(|\Delta_{J^*}|_1 - \overline{\beta}|\Psi\Delta|_\infty) + (1 + c_0)\overline{\beta}|\Psi\Delta|_\infty \\ &\leqslant (1 + c_0)\alpha_s(X)|\Delta|_1 + (1 + c_0)\overline{\beta}|\Psi\Delta|_\infty. \end{aligned} \tag{3.26}$$

Combining (3.24) and (3.26) yields the results. □

Chapter 4

Sup-norm convergence rate of the Lasso under the Irrepresentable Condition

We derive the l_∞ convergence rate and the sign concentration property of the Lasso in a high-dimensional Gaussian linear regression model under the Irrepresentable Condition on the Gram matrix of the design.

4.1 Introduction

In this chapter, we derive some properties of the Lasso estimator under the Strong Irrepresentable Condition $\mathbf{IC}(s, 1-\gamma)$ on the Gram matrix of the design for some $\gamma > 0$. We prove in particular that the solution of the Lasso minimization problem is unique with overwhelming probability and give its explicit form. We also establish the sup-norm convergence rate and variable selection properties of the Lasso estimator under this condition.

Consider the linear regression model

$$Y = X\theta^* + W, \tag{4.1}$$

where X is a $n \times M$ deterministic matrix, $\theta^* \in \mathbb{R}^M$ and $W = (W_1, \ldots, W_n)^T$ is a zero-mean random vector such that $\mathbb{E}[W_i^2] \leqslant \sigma^2$, $1 \leqslant i \leqslant n$ for some $\sigma^2 > 0$. For any $\theta \in \mathbb{R}^M$, define $J(\theta) = \{j \ : \ \theta_j \neq 0\}$. Let $M(\theta) = |J(\theta)|$ be the cardinality of $J(\theta)$ and $\overrightarrow{\text{sign}}(\theta) = (\text{sign}(\theta_1), \ldots, \text{sign}(\theta_M))^T$ where

$$\text{sign}(t) = \begin{cases} 1 & \text{if } t > 0, \\ 0 & \text{if } t = 0, \\ -1 & \text{if } t < 0. \end{cases}$$

For any $j \in \{1, \ldots, M\}$, denote by X_j the j-th column of X. For any vector $\theta \in \mathbb{R}^M$ and any subset J of $\{1, \ldots, M\}$, we denote by θ_J the vector in $\mathbb{R}^M$ which has the same coordinates as θ on J and zero coordinates on the complement J^c of J, by $\overline{\theta}_J$ the vector in $\mathbb{R}^J$ obtained by keeping only the components θ_j with index $j \in J$ and by X_J the $n \times |J|$ sub-matrix of X obtained by keeping the columns of X with index $j \in J$. For any integers $1 \leqslant d, p < \infty$ and $z = (z_1, \ldots, z_d) \in \mathbb{R}^d$, the l_p norm of the vector z is denoted by $|z|_p \stackrel{\Delta}{=} \left(\sum_{j=1}^d |z_j|^p\right)^{1/p}$, and $|z|_\infty \stackrel{\Delta}{=} \max_{1 \leqslant j \leqslant d} |z_j|$.

Recall that in the case $M > n$, if a vector $\theta^* = \theta^0$ satisfies (4.1), then there exists an affine space $\Theta^* = \{\theta^* \ : \ X\theta^* = X\theta^0\}$ of dimension larger than $M - n$ of vectors satisfying (4.1). However, we proved in Theorem 3.3 in Section 3.3 that if the Gram matrix of the design satisfies Assumption $\mathbf{IC}(s, 1)$, then the s-sparse vector θ^* satisfying (4.1) is the unique solution of the l_1-minimization problem (3.17). This settles the identifiability problem when $M > n$ since $\mathbf{IC}(s, 1-\gamma)$ is more restrictive than $\mathbf{IC}(s, 1)$.

The Lasso estimators $\hat{\theta}^L$ solve the minimization problem

$$\min_{\theta \in \mathbb{R}^M} \frac{1}{n}|Y - X\theta|_2^2 + 2r|\theta|_1, \tag{4.2}$$

where $r > 0$ is a constant. A convenient choice in our context will be $r = A\sigma\sqrt{(\log M)/n}$, for some $A > 0$. We denote by $\hat{\Theta}^L$ the set of solutions to the Lasso minimization problem (4.2).

Define $\Phi(\theta) = \frac{1}{n}|Y - X\theta|_2^2 + 2r|\theta|_1$. The necessary and sufficient condition for $\hat{\theta}^L$ to minimize Φ is that the zero vector in $\mathbb{R}^M$ belongs to the subdifferential of Φ at point $\hat{\theta}^L$, i.e.,

$$\begin{cases} \frac{1}{n}(X^T(Y - X\hat{\theta}^L))_j = \text{sign}(\hat{\theta}_j^L)r & \text{if } \hat{\theta}_j \neq 0, \\ \left|\frac{1}{n}(X^T(Y - X\hat{\theta}^L))_j\right| \leqslant r & \text{if } \hat{\theta}_j^L = 0. \end{cases} \tag{4.3}$$

The Lasso estimator is unique if $M < n$, since in this case $\Phi(\theta)$ is strongly convex. However, for $M > n$ it is not necessarily unique. However when $M > n$, we show that the Lasso solution is unique with probability close to 1 under the Strong Irrepresentable Condition $\mathbf{IC}(s, 1-\gamma)$.

Now we state the assumptions on our model. The first assumption concerns the noise variables.

Assumption 4.1. *The random variables $W_1, \ldots, W_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$.*

Define the Gram matrix of the design by

$$\Psi \stackrel{\triangle}{=} \frac{1}{n}X^T X.$$

For any subsets J, J' of $\{1, \ldots, M\}$, denote by $\Psi_{J,J}$ the following submatrix of Ψ

$$\Psi_{J,J'} \stackrel{\triangle}{=} (\Psi_{i,j})_{i \in J, j \in J'}.$$

We assume w.l.o.g. that

$$\Psi_{j,j} = 1, \, \forall 1 \leqslant j \leqslant M.$$

If this is not the case we can prove exactly the same results considering the following Lasso minimization problem

$$\min_{\theta \in \mathbb{R}^M} \frac{1}{n}|Y - X\theta|_2^2 + r\sum_{j=1}^{M} \Psi_{j,j}^{1/2}|\theta_j|. \tag{4.4}$$

We recall below the Irrepresentable Condition $\mathbf{IC}(s, \eta)$ on Ψ.

Assumption 4.2. *Let $s \geqslant 1$ be an integer and fix an arbitrary $\eta \in (0, 1/2)$. The $n \times M$ matrix X satisfies the Irrepresentable Condition $\boldsymbol{IC}(s, \eta)$ if for any subset I of $\{1, \ldots, M\}$ with $|I| \leqslant s$ we have:*

– *the matrix X_I has trivial kernel,*

– *for any vector* z *in* $\{-1,1\}^{|I|}$, *we have*

$$|X_{I^c}^T X_I (X_I^T X_I)^{-1} z|_\infty < \eta. \tag{4.5}$$

In the presence of noise, we need the following assumption on the nonzero components of θ^*. Set $J^* = J(\theta^*)$ and $\Psi_* = \Psi_{J^*,J^*}$.

Assumption 4.3. *We have*

$$\rho \stackrel{\Delta}{=} \min_{j \in J(\theta^*)} |\theta_j^*| > r \max_{j \in J^*} \left(\sqrt{(\Psi_*^{-1})_{j,j}} + |(\Psi_*^{-1} \overrightarrow{\text{sign}}(\overline{\theta^*}_{J^*}))_j| \right).$$

This assumption says that the nonzero components of θ^* cannot be arbitrarily small. This is needed to prove the l_∞-estimation rate or the sign concentration property. We can find similar assumptions on ρ in the work on sign consistency of the Lasso estimator mentioned above. More precisely, the lower bound on ρ is of the order $s^{1/4} r^{1/2}$ in [82], $n^{-\delta/2}$ with $0 < \delta < 1$ in [104, 116], $\sqrt{(\log Mn)/n}$ in [12], $\sqrt{s}r$ in [114] and r in [71]. Note that in our assumption, the quantity $\max_{j \in J^*} |(\Psi_*^{-1} \overrightarrow{\text{sign}}(\overline{\theta^*}_{J^*}))_j|$ can be bounded from above by an absolute constant or can be of the order s, depending on the considered Gram matrix Ψ. In Section 4.2 we prove some preliminary results. In Section 4.3 we derive the uniqueness and the l_∞ estimation rate of the Lasso under the irrepresentable condition and we compare our result to the existing results in the literature.

4.2 Preliminary results

Let $\hat{\theta}^0 \in \mathbb{R}^{M(\theta^*)}$ be the least squares estimator of $\overline{\theta^*}_{J^*}$ if the set of nonzero components of θ^* were known in advance. Since we assumed that $M(\theta^*) \leqslant s$, the first point of Assumption 4.2 ensures that the matrix Ψ_* is positive definite. Thus

$$\hat{\theta}^0 = \frac{1}{n} \Psi_*^{-1} X_{J^*}^T Y \sim \mathcal{N}\left(\overline{\theta^*}_{J^*}, \frac{\sigma^2}{n} \Psi_*^{-1} \right).$$

Consider the vector $\tilde{\theta}^0 \in \mathbb{R}^M$ such that

$$\tilde{\theta}_j^0 = 0, \ \forall j \notin J^*, \tag{4.6}$$

and

$$\overline{\tilde{\theta}^0}_{J^*} = \hat{\theta}^0 - r \Psi_*^{-1} \overrightarrow{\text{sign}}(\overline{\theta^*}_{J^*}). \tag{4.7}$$

Lemma 4.1. *Assume that the diagonal elements of Ψ are equal to 1 and that $M(\theta^*) \leqslant s$. Fix arbitrary $A > 0$ and $\eta > 0$. Define $r = A\sigma\sqrt{(\log M)/n}$ and the event*

$$\mathcal{A} = \bigcap_{j \in J^{*c} \,:\, \Psi_{j,J^*} \neq 0} \left\{ \frac{1}{\sqrt{\Psi_{j,J^*}\Psi_*^{-1}\Psi_{J^*,j}}} |\Psi_{j,J^*}(\hat{\theta}^0 - \overline{\theta^*}_{J^*})| \leqslant \eta r \right\} \bigcap_{j \in J^*} \left\{ \frac{1}{\sqrt{(\Psi_*^{-1})_{j,j}}} |\hat{\theta}_j^0 - \theta_j^*| \leqslant r \right\} \bigcap \left\{ \frac{1}{n} |X^T W|_\infty \leqslant \eta r \right\}. \tag{4.8}$$

Then, we have $\mathbb{P}(\mathcal{A}) \geqslant 1 - 2M^{1-\frac{(\eta A)^2}{2}} - sM^{-\frac{A^2}{2}}$.

Proof. Define the random variables $Z_j = n^{-1}\sum_{i=1}^n X_{i,j}W_i$, $1 \leqslant j \leqslant M$. We have that $Z_j \sim \mathcal{N}(0, \sigma^2/n)$, $1 \leqslant j \leqslant M$. By definition of $\hat{\theta}^0$, we get for any $j \in J^*$

$$Z_j' = \frac{1}{\sqrt{(\Psi_*^{-1})_{j,j}}}(\hat{\theta}_j^0 - \theta_j^*) \sim \mathcal{N}\left(0, \sigma^2/n\right),$$

and for any $j \in J^{*c}$ such that $\Psi_{j,J^*} \neq 0$

$$Z_j'' = \frac{1}{\sqrt{\Psi_{j,J^*}\Psi_*^{-1}\Psi_{J^*,j}}} \Psi_{j,J^*}(\hat{\theta}^0 - \theta^*) \sim \mathcal{N}\left(0, \sigma^2/n\right),$$

Then

$$\mathcal{A} = \bigcap_{j \in J^{*c} : \Psi_{j,J^*} \neq 0} \{|Z_j''| \leqslant \eta r\} \cap \bigcap_{j \in J^*} \{|Z_j'| \leqslant r\} \cap \bigcap_{j=1}^M \{|Z_j| \leqslant \eta r\}.$$

Set $Z, Z', Z'' \sim \mathcal{N}(0, \frac{\sigma^2}{n})$. Standard inequalities on the tail of Gaussian variables yield

$$\begin{aligned} P(\mathcal{A}^c) &\leqslant MP(|Z''| \geqslant \eta r) + M(\theta^*)P(|Z'| \geqslant r) + MP(|Z| \geqslant \eta r), \\ &\leqslant 2M \exp\left(-\frac{n}{2\sigma^2}(\eta r)^2\right) + M(\theta^*) \exp\left(-\frac{n}{2\sigma^2} r^2\right) \\ &\leqslant 2M^{1-\frac{(\eta A)^2}{2}} + sM^{-\frac{A^2}{2}}. \end{aligned}$$

□

We now state that $\tilde{\theta}^0$ satisfies a slightly stronger condition than (4.3) on the event $\mathcal{A}$.

Lemma 4.2. *Let $s \geqslant 1$ be an integer. Fix an arbitrary $\eta \in (0, 1/2)$. Let the event $\mathcal{A}$ be defined in (4.8). Let Assumptions 4.1, 4.3 be satisfied. Assume that X satisfies the condition $\boldsymbol{IC}(s, 1-2\eta)$. If $M(\theta^*) \leqslant s$, then the vector $\tilde{\theta}^0$ defined above satisfies the following relations on the event $\mathcal{A}$*

$$\begin{cases} \frac{1}{n}X_j^T(Y - X\tilde{\theta}^0) = r\mathrm{sign}(\tilde{\theta}_j^0), & \text{if } \tilde{\theta}_j^0 \neq 0, \\ |\frac{1}{n}X_j^T(Y - X\tilde{\theta}^0)| < r, & \text{if } \tilde{\theta}_j^0 = 0. \end{cases} \tag{4.9}$$

Proof. Since $\hat{\theta}^0$ is the least squares estimator of $\overline{\theta^*}_{J^*}$, we have

$$\frac{1}{n}X_{J^*}^T(Y - X_{J^*}\hat{\theta}^0) = \vec{0}.$$

Thus,

$$\frac{1}{n}X_{J^*}^T(Y - X\tilde{\theta}^0) = r\overrightarrow{\text{sign}}(\overline{\theta^*}_{J^*}).$$

We now prove that on the event $\mathcal{A}$, $\text{sign}(\theta_j^*) = \text{sign}(\tilde{\theta}_j^0)$ for any $j \in J^*$. Clearly, we have on the event $\mathcal{A}$

$$|\hat{\theta}_j^0 - \theta_j^*| \leqslant r\sqrt{(\Psi_*^{-1})_{j,j}},$$

for any $j \in J^*$. This yields by definition of $\tilde{\theta}^0$

$$|\tilde{\theta}_j^0 - \theta_j^*| \leqslant r\left(\sqrt{(\Psi_*^{-1})_{j,j}} + \left|(\Psi_*^{-1}\overrightarrow{\text{sign}}(\overline{\theta^*}_{J^*}))_j\right|\right), \forall j \in J^*. \tag{4.10}$$

Then, Assumption 4.3 implies that $\text{sign}(\tilde{\theta}_j^0) = \text{sign}(\theta_j^*)$ for any $j \in J^*$.

Next, we have for any $j \notin J^*$

$$\begin{aligned}\frac{1}{n}X_j^T(Y - X_{J^*}\overline{\tilde{\theta}^0}_{J^*}) &= \frac{1}{n}X_j^T W + \Psi_{j,J^*}(\overline{\theta^*}_{J^*} - \overline{\tilde{\theta}^0}_{J^*}) \\ &= \frac{1}{n}X_j^T W + \Psi_{j,J^*}(\overline{\theta^*}_{J^*} - \hat{\theta}^0) + r\Psi_{j,J^*}\Psi_*^{-1}\overrightarrow{\text{sign}}(\overline{\theta^*}_{J^*}). \end{aligned} \tag{4.11}$$

On the event $\mathcal{A}$, we have

$$|\Psi_{j,J^*}(\overline{\theta^*}_{J^*} - \hat{\theta}^0)| \leqslant (\Psi_{j,J^*}\Psi_*^{-1}\Psi_{J^*,j})^{1/2}\eta r, \tag{4.12}$$

and

$$\frac{1}{n}|X_j^T W| \leqslant \eta r. \tag{4.13}$$

Combining equations (4.11)-(4.13) with Assumption $\mathbf{IC}(s, 1-2\eta)$ we obtain on the event $\mathcal{A}$

$$\frac{1}{n}\left|X_j^T(Y - X_{J^*}\overline{\tilde{\theta}^0}_{J^*})\right| \leqslant \eta r + (\Psi_{j,J^*}\Psi_*^{-1}\Psi_{J^*j})^{1/2}\eta r + r(1-2\eta),\ \forall j \in J^{*c}.$$

Next, we have

$$\Psi_{j,J^*}\Psi_*^{-1}\Psi_{J^*j} \leqslant |\Psi_{j,J^*}\Psi_*^{-1}|_1|\Psi_{J^*,j}|_\infty \leqslant |\Psi_{j,J^*}\Psi_*^{-1}|_1,$$

since $|\Psi_{j,k}| \leqslant 1$ for any $j \neq k$. Set $u_j = \Psi_{j,J^*}\Psi_*^{-1}$ in $\mathbb{R}^{|J^*|}$. Consider the sign vector $z = \overrightarrow{\text{sign}}(u_j)$ in $\mathbb{R}^{|J^*|}$. We have

$$|u_j|_1 = \Psi_{j,J^*}(\Psi^*)^{-1}z < 1 - 2\eta,$$

since Ψ satisfies Assumption $\mathbf{IC}(s, 1-2\eta)$. The three above displays yield

$$\frac{1}{n}\left|X_j^T(Y - X_{J^*}\overline{\tilde{\theta}^0}_{J^*})\right| \leqslant r(\eta + (1-2\eta)^{1/2}\eta + (1-2\eta)) < r,\ \forall j \in J^{*c}.$$

□

We will need the following simple lemma [95].

Lemma 4.3. *Let $\hat{\theta}^L$ be a Lasso solution such that*

$$\frac{1}{n}|X_j^T(Y - X\hat{\theta}^L)| < r, \ \ \forall j \in \hat{S}^c,$$

where $\hat{S} = \{j : \hat{\theta}_j^L \neq 0\}$. Then for any other solution $\tilde{\theta} \in \hat{\Theta}^L$ we have

$$\tilde{\theta}_{\hat{S}^c} = \hat{\theta}^L_{\hat{S}^c} = 0, \ \ \textit{and} \ \ X(\tilde{\theta} - \hat{\theta}^L) = 0.$$

Proof. For any $\theta \in \mathbb{R}^M$ define

$$\Phi(\theta) = \frac{1}{n}|Y - X\theta|_2^2 + 2r|\theta|_1.$$

For any $h \in \mathbb{R}^M$ we have

$$\begin{aligned}\Phi(\hat{\theta}^L + h) = \frac{1}{n}|Y - X\hat{\theta}^L|_2^2 - \frac{2}{n}h^T X^T(Y - X\hat{\theta}^L) + \frac{1}{n}|Xh|_2^2 \\ + 2r|\hat{\theta}^L_{\hat{S}} + h_{\hat{S}}|_1 + 2r|\hat{\theta}^L_{\hat{S}^c} + h_{\hat{S}^c}|_1.\end{aligned}$$

Note that, by convexity of $|\cdot|_1$,

$$|\hat{\theta}^L_{\hat{S}} + h_{\hat{S}}|_1 \geqslant |\hat{\theta}^L_{\hat{S}}|_1 + \overrightarrow{\text{sign}}(\hat{\theta}^L_{\hat{S}})^T h_{\hat{S}}.$$

This yields

$$\begin{aligned}\Phi(\hat{\theta}^L + h) \geqslant &\Phi(\hat{\theta}^L) + 2r|h_{\hat{S}^c}|_1 \\ &+ 2r\overrightarrow{\text{sign}}(\hat{\theta}^L_{\hat{S}})^T h_{\hat{S}} - \frac{2}{n}h^T X^T(Y - X\hat{\theta}^L) + \frac{1}{n}|Xh|_2^2.\end{aligned}$$

Since $\hat{\theta}^L$ is a minimizer of Φ, it satisfies

$$\frac{1}{n}X^T_{(j)}(Y - X\hat{\theta}^L) = r\text{sign}(\hat{\theta}^L_j), \ \ \forall j \in \hat{S}.$$

We also have by our assumption that

$$\left|\frac{1}{n}X_j^T(Y - X\hat{\theta}^L)\right| < r, \ \ \forall j \in \hat{S}^c.$$

This yields

$$\begin{aligned}\frac{2}{n}h^T X^T(Y - X\hat{\theta}^L) &= \frac{2}{n}\sum_{j\in\hat{S}} h_j X_j^T(Y - X\hat{\theta}^L) + \frac{2}{n}\sum_{j\in\hat{S}^c} h_j X_j^T(Y - X\hat{\theta}^L) \\ &= 2r\sum_{j\in\hat{S}} h_j \text{sign}(\hat{\theta}^L_j) + 2r\sum_{j\in\hat{S}^c} h_j\epsilon_j,\end{aligned}$$

where the $\epsilon_j = \frac{1}{n} X_j^T (Y - X\hat{\theta}^L)$ are such that $|\epsilon_j| < 1$ for all $j \in \hat{S}^c$.

Thus

$$\Phi(\hat{\theta}^L + h) = \Phi(\hat{\theta}^L) + 2r|h_{\hat{S}^c}|_1 - 2r \sum_{j \in \hat{S}^c} h_j \epsilon_j + \frac{1}{n}|Xh|_2^2.$$

Now let $\tilde{\theta}$ be a minimizer of Φ distinct from $\hat{\theta}^L$. Set $h = \tilde{\theta} - \hat{\theta}^L$. The above display yields the result since $|\epsilon_j| < 1$ for all $j \in \hat{S}^c$. □

4.3 Sup-norm estimation and variable selection with the Lasso

We state now our main result of the chapter concerning the Lasso estimator.

Theorem 4.1. *Let $s \geqslant 1$ be an integer. Fix an arbitrary $\eta \in (0, 1/2)$. Take $r = A\sigma\sqrt{(\log M)/n}$ and $A > \sqrt{2}/\eta$. Let Assumptions 4.1, 4.3 be satisfied. Assume that X satisfies the condition $\boldsymbol{IC}(s, 1-2\eta)$. If $M(\theta^*) \leqslant s$, then we have, with probability at least $1 - 2M^{1-\frac{(\eta A)^2}{2}} - sM^{-\frac{A^2}{2}}$:*

1. *the Lasso solution $\hat{\theta}^L$ is unique and equal to $\tilde{\theta}^0$ defined in (4.6)-(4.7),*
2. $$|\hat{\theta}^L - \theta^*|_\infty \leqslant r \max_{j \in J^*} \left(\sqrt{(\Psi^{*-1})_{j,j}} + \left| (\Psi_*^{-1} \overrightarrow{\operatorname{sign}}(\overline{\theta^*}_{J^*}))_j \right| \right), \tag{4.14}$$
3. $$\overrightarrow{\operatorname{sign}}(\hat{\theta}^L) = \overrightarrow{\operatorname{sign}}(\theta^*).$$

Proof. We prove the first point. Lemma 4.2 guarantees that $\tilde{\theta}^0$ satisfies the necessary and sufficient conditions to be a Lasso solution on the event $\mathcal{A}$. Assume there exists a solution $\hat{\theta}^1$ in $\hat{\Theta}^L$ distinct from $\tilde{\theta}^0$. Lemma 4.3 implies that on the event $\mathcal{A}$

$$J(\hat{\theta}^1) \subset J(\tilde{\theta}^0) = J^*, \text{ and } X\tilde{\theta}^0 = X\hat{\theta}^1.$$

Then, since $X\tilde{\theta}^0 = X\hat{\theta}^1$, $|J^*| \leqslant s$ and X_{J^*} has trivial kernel, by the first point of Assumption 4.2 we have $\tilde{\theta}^0 = \hat{\theta}^1$.

The second point is immediate by definition of $\tilde{\theta}^0$ and the inequality (4.10). We apply Lemma 4.1 to conclude. □

We now comment on some results similar to Theorem 4.1. Dossal [35] established the uniqueness of the Lasso solution in the noiseless case under Assumption $\mathbf{IC}(s, 1)$ and some additional conditions on the design matrix X and $\overrightarrow{\operatorname{sign}}(\overline{\theta^*}_{J^*})$, which can be compared to 1 of Theorem 4.1. Wainwright [103] derived recently a result which is quite similar to

ours. However, unlike our results, [103] did not give the explicit representation of the Lasso estimator. Note also that the bound in (4.14) is better than that obtained in Theorem 1 of [103]. Indeed, the leading term in (4.14) is $\max_{j\in J^*}\left|\sum_{k\in J^*}(\Psi_*^{-1})_{j,k}\mathrm{sign}(\theta_k^*)\right|$ as compared to the leading term in the bound obtained in [103]:

$$|||\Psi_*^{-1}|||_\infty = \max_{j\in J^*}\sum_{k\in J^*}|(\Psi_*^{-1})_{j,k}|,$$

where $(\Psi_*^{-1})_{j,k}$ are the elements of the matrix Ψ_*^{-1}. Note that the result of the type (4.14) is stated in [103] without the condition that the nonzero components of θ^* are sufficiently large. However, [103] overlooked that this condition is actually needed. Indeed, [103] built the solution estimator, that is the estimator satisfying (4.9), under the following condition (see Lemma 3 (b)):

$$\mathrm{sign}(\theta_j^* + \Delta_j) = \mathrm{sign}(\theta_j^*), \quad \forall j \in J(\theta^*),$$

where the $\Delta_j \in \mathbb{R}$ are similar to the quantities appearing on the right-hand-side of (4.10). Thus, the condition in the above display is satisfied if and only if the nonzero components of θ^* are large enough. The restriction on $\rho = \min_{j\in J(\theta^*)}|\theta_j^*|$ to be large enough is inevitable when we try to exploit the optimality conditions (4.9) to establish the sup-norm estimation result. However, we would like to emphasize that we do not need any condition on the nonzero components of θ^* to obtain the same sup-norm estimation result under the mutual coherence condition for the Lasso and the Dantzig Selector, see Theorem 2.1.

In Chapter 2, we proved that the Lasso and Dantzig estimators achieve the optimal rate of sup-norm convergence under a mutual coherence assumption on the Gram matrix of the design. We say that a sup-norm estimation rate is optimal if it is of the form $\alpha\sigma\sqrt{\frac{\log M}{n}}$ where $\alpha > 0$ is an absolute constant as in the case of gaussian sequence model ($n = M$ and $\Psi = I_M$ where I_M denotes the $M \times M$ identity matrix). Define

$$d^* = |\Psi_*^{-1}\overrightarrow{\mathrm{sign}}(\overline{\theta^*}_{J^*})|_\infty. \tag{4.15}$$

In view of Theorem 4.1, we see that d^* is the important quantity to determine whether the obtained sup-norm convergence rate is optimal. Note that this quantity can be linear in s for some vectors θ^* and some matrices X. In this case, we do not get the optimal rate.

Assume now that Ψ satisfies a mutual coherence condition $\max_{i\neq j}|\Psi_{i,j}| \leqslant \frac{1}{cs}$ for some $c > 0$. Then, similarly as we proved (3.10), we get that

$$\max_{i\neq j}|(\Psi_*^{-1})_{i,j}| \leqslant \frac{1}{s(c-1)}.$$

This yields that Ψ_*^{-1} also satisfies a mutual coherence condition and that

$$d^* = |\Psi_*^{-1}\overrightarrow{\text{sign}}(\overline{\theta^*}_{J^*})|_\infty \leqslant \frac{c}{c-1}.$$

Note that if we take $c = 7\alpha$ with $\alpha > 0$ (we have $7 = 1 + 2c_0$ with $c_0 = 3$ since we consider the Lasso estimator), then condition $\mathbf{IC}\,(s, 1-2\eta)$ is satisfied for

$$\alpha \geqslant \frac{3-4\eta}{7(1-2\eta)}.$$

For $\eta \leqslant 1/4$, we have the condition $\alpha \geqslant 4/7$. Thus, we improve upon the results of Chapter 2 where the restriction $\alpha > 1$ was imposed in Assumption 2.2.

Note also that in this case, the restriction on ρ takes the form $\rho > Cr$ where $C > 0$ is a constant independent of s. This is coherent with the results obtained directly under the mutual coherence assumption in Chapter 2.

Chapter 5

Taking Advantage of Sparsity in Multi-Task Learning

This chapter contains the results of the article [74], written in collaboration with M. Pontil, A.B. Tsybakov and S.A. van de Geer.

We study the problem of estimating multiple linear regression equations for the purpose of both prediction and variable selection. Following recent work on multi-task learning [1], we assume that the regression vectors share the same sparsity pattern. This means that the set of relevant predictor variables is the same across the different equations. This assumption leads us to consider the Group Lasso as a candidate estimation method. We show that this estimator enjoys nice sparsity oracle inequalities and variable selection properties. The results hold under a certain restricted eigenvalue condition and a coherence condition on the design matrix, which naturally extend recent work in [6, 71]. In particular, in the multi-task learning scenario, in which the number of tasks can grow, we are able to remove completely the effect of the number of predictor variables in the bounds. Finally, we show how our results can be extended to more general noise distributions, of which we only require the fourth moment to be finite.

5.1 Introduction

We study the problem of estimating multiple regression equations under sparsity assumptions on the underlying regression coefficients. More precisely, we consider multiple Gaussian regression models,

$$\begin{aligned} y_1 &= X_1\beta_1^* + W_1 \\ y_2 &= X_2\beta_2^* + W_2 \\ &\vdots \\ y_T &= X_T\beta_T^* + W_T \end{aligned} \tag{5.1}$$

where, for each $t = 1, \ldots, T$, we let X_t be a prescribed $n \times M$ design matrix, β_t^* the unknown vector of regression coefficients and y_t an n-dimensional vector of observations. We assume that $W_1, \ldots, W_T$ are *i.i.d.* zero mean random vectors.

We are interested in estimation methods which work well even when the number of parameters in each equation is much larger than the number of observations, that is, $M \gg n$. This situation may arise in many practical applications in which the predictor variables are inherently high dimensional, or it may be "costly" to observe response variables, due to difficult experimental procedures, see, for example [1] for a discussion.

Examples in which this estimation problem is relevant range from multi-task learning [1, 19, 76, 86] and conjoint analysis (see, for example, [38, 67] and references therein) to longitudinal data analysis [29] as well as the analysis of panel data [53, 108], among others. In particular, multi-task learning provides a main motivation for our study. In that setting each regression equation corresponds to a different learning task (the classification case can be treated similarly); in addition to the requirement that $M \gg n$, we also allow for the number of tasks T to be much larger than n. Following [1] we assume that there are only few common important variables which are shared by the tasks. A general goal of this chapter is to study the implications of this assumption from a statistical learning view point, in particular, to quantify the advantage provided by the large number of tasks to learn the underlying vectors $\beta_1^*, \ldots, \beta_T^*$ as well as to select common variables shared by the tasks.

Our study pertains and draws substantial ideas from the recently developed area of compressed sensing and sparse estimation (or sparse recovery), see [6, 16, 34] and references therein. A central problem studied therein is that of estimating the parameters of a (single) Gaussian regression model. Here, the term "sparse" means that most of the components of the underlying M-dimensional regression vector are equal to zero. A main motivation for sparse estimation comes from the observation that in many practical applications M is much larger than the number n of observations but the underlying model is sparse, see [16, 34] and

references therein. Under this circumstance ordinary least squares will not work. A more appropriate method for sparse estimation is the ℓ_1-norm penalized least squares method, which is commonly referred to as the Lasso method. In fact, it has been recently shown by different authors, under different conditions on the design matrix, that the Lasso satisfies sparsity oracle inequalities, see [6, 15, 14, 99] and references therein. Closest to our study in this chapter is [6], which relies upon a Restricted Eigenvalue (RE) assumption. The results of these works make it possible to estimate the parameter β even in the so-called "*p much larger than n*" regime (in our notation, the number of predictor variables p corresponds to MT).

In this chapter, we assume that the vectors $\beta_1^*, \ldots, \beta_T^*$ are not only sparse but also have their sparsity patterns included in the same set of small cardinality s. In other words, the response variable associated with each equation in (5.1) depends only on some members of a small subset of the corresponding predictor variables, which is preserved across the different equations. This assumption, that we further refer to as *structured sparsity assumption*, is motivated by some recent work on multi-task learning [1]. It naturally leads to an extension of the Lasso method, the so-called Group Lasso [112], in which the error term is the average residual error across the different equations and the penalty term is a mixed $(2,1)$-norm. The structured sparsity assumption induces a relation between the responses and, as we shall see, can be used to improve estimation.

The chapter is organized as follows. In Section 5.2 we define the estimation method and comment on previous related work. In Section 5.3 we study the oracle properties of this estimator when the errors W_t are Gaussian. Our main results concern upper bounds on the prediction error and the distance between the estimator and the true regression vector β^*. Specifically, Theorem 5.1 establishes that under the above structured sparsity assumption on β^*, the prediction error is essentially of the order of s/n. In particular, in the multi-task learning scenario, in which T can grow, we are able to remove completely the effect of the number of predictor variables in the bounds. Next, in Section 5.4, under a stronger condition on the design matrices, we describe a simple modification of our method and show that it selects the correct sparsity pattern with an overwhelming probability (Theorem 5.2). We also find the rates of convergence of the estimators for mixed $(2,1)$-norms with $1 \leq p \leq \infty$ (Corollary 5.1). The techniques of proofs build upon and extend those of [6] and [71]. Finally, in Section 5.5 we discuss how our results can be extended to more general noise distributions, of which we only require the fourth moment to be finite. Specifically, we prove a generalization of Nemirovski's moment inequality which is interesting by itself.

5.2 Method and related work

In this section we first introduce some notation and then describe the estimation method which we analyze in the chapter. As stated above, our goal is to estimate T linear regression functions identified by the parameters $\beta_1^*, \ldots, \beta_T^* \in \mathbb{R}^M$. We may write the model (5.1) in compact notation as

$$y = X\beta^* + W \tag{5.2}$$

where y and W are the nT-dimensional random vectors formed by stacking the vectors $y_1, \ldots, y_T$ and the vectors $W_1, \ldots, W_T$, respectively. Likewise β^* denotes the vector obtained by stacking the regression vectors $\beta_1^*, \ldots, \beta_T^*$. Unless otherwise specified, all vectors are meant to be column vectors. Thus, for every $t \in \mathbb{N}_T$, we write $y_t = (y_{ti} : i \in \mathbb{N}_n)^\top$ and $W_t = (W_{ti} : i \in \mathbb{N}_n)^\top$, where, hereafter, for every positive integer k, we let $\mathbb{N}_k$ be the set of integers from 1 and up to k. The $nT \times MT$ block diagonal design matrix X has its t-th block formed by the $n \times M$ matrix X_t. We let $x_{t1}^\top, \ldots, x_{tn}^\top$ be the row vectors forming X_t and $(x_{ti})_j$ the j-th component of the vector x_{ti}. Throughout the chapter we assume that x_{ti} are deterministic.

For every $\beta \in \mathbb{R}^{MT}$ we introduce $(\beta)^j \equiv \beta^j = (\beta_{tj} : t \in \mathbb{N}_T)^\top$, that is, the vector formed by the coefficients corresponding to the j-th variable. For every $1 \leq p < \infty$ we define the mixed $(2,p)$-norm of β as

$$\|\beta\|_{2,p} = \left(\sum_{j=1}^{M} \left(\sum_{t=1}^{T} \beta_{tj}^2 \right)^{\frac{p}{2}} \right)^{\frac{1}{p}} = \left(\sum_{j=1}^{M} \|\beta^j\|^p \right)^{\frac{1}{p}}$$

and the $(2,\infty)$-norm of β as

$$\|\beta\|_{2,\infty} = \max_{1 \leq j \leq M} \|\beta^j\|,$$

where $\|\cdot\|$ is the standard Euclidean norm.

If $J \subseteq \mathbb{N}_M$ we let $\beta_J \in \mathbb{R}^{MT}$ be the vector formed by stacking the vectors $(\beta^j I\{j \in J\} : j \in \mathbb{N}_M)$, where $I\{\cdot\}$ denotes the indicator function. Finally we set $J(\beta) = \{j : \beta^j \neq 0,\ j \in \mathbb{N}_M\}$ and $M(\beta) = |J(\beta)|$ where $|J|$ denotes the cardinality of set $J \subset \{1, \ldots, M\}$. The set $J(\beta)$ contains the indices of the relevant variables shared by the vectors $\beta_1, \ldots, \beta_T$ and the number $M(\beta)$ quantifies the level of structured sparsity across those vectors.

We have now accumulated the sufficient information to introduce the estimation method. We define the empirical residual error

$$\hat{S}(\beta) = \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} (x_{ti}^\top \beta_t - y_{ti})^2 = \frac{1}{nT} \|X\beta - y\|^2$$

and, for every $\lambda > 0$, we let our estimator $\hat{\beta}$ be a solution of the optimization problem [1]

$$\min_{\beta} \hat{S}(\beta) + 2\lambda \|\beta\|_{2,1}. \tag{5.3}$$

In order to study the statistical properties of this estimator, it is useful to derive the optimality condition for a solution of the problem (5.3). Since the objective function in (5.3) is convex, $\hat{\beta}$ is a solution of (5.3) if and only if 0 (the MT-dimensional zero vector) belongs to the subdifferential of the objective function. In turn, this condition is equivalent to the requirement that

$$-\nabla \hat{S}(\hat{\beta}) \in 2\lambda \partial \left(\sum_{j=1}^{M} \|\hat{\beta}^j\| \right),$$

where ∂ denotes the subdifferential (see, for example, [10] for more information on convex analysis). Note that

$$\partial \left(\sum_{j=1}^{M} \|\beta^j\| \right) = \left\{ \theta \in \mathbb{R}^{MT} : \theta^j = \frac{\beta^j}{\|\beta^j\|} \text{ if } \beta^j \neq 0,\ \|\theta^j\| \leq 1, \text{ if } \beta^j = 0,\ j \in \mathbb{N}_M \right\}.$$

Thus, $\hat{\beta}$ is a solution of (5.3) if and only if

$$\frac{1}{nT}(X^\top (y - X\hat{\beta}))^j = \lambda \frac{\hat{\beta}^j}{\|\hat{\beta}^j\|}, \quad \text{if } \hat{\beta}^j \neq 0 \tag{5.4}$$

$$\frac{1}{nT}\|(X^\top (y - X\hat{\beta}))^j\| \leq \lambda, \quad \text{if } \hat{\beta}^j = 0. \tag{5.5}$$

We now comment on previous related work. Our estimator is a special case of the Group Lasso estimator [112]. Several papers analyzing statistical properties of the Group Lasso appeared quite recently [4, 22, 54, 64, 78, 79, 83, 88]. Most of them are focused on the Group Lasso for additive models [54, 64, 79, 88] or generalized linear models [78]. Special choice of groups is studied in [22]. Discussion of the Group Lasso in a relatively general setting is given by Bach [4] and Nardi and Rinaldo [83]. Bach [4] assumes that the predictors x_{ti} are random with a positive definite covariance matrix and proves results on consistent selection of sparsity pattern $J(\beta^*)$ when the dimension of the model ($p = MT$ in our case) is fixed and $n \to \infty$. Nardi and Rinaldo [83] consider a setting that covers ours and address the issue of sparsity oracle inequalities in the spirit of [6]. However, their bounds are too coarse (see comments in Section 5.3 below). Obozinski et al. [25] consider the case where all the matrices X_i are the same and all their rows are independent Gaussian random vectors with the same covariance matrix. They show that the resulting estimator achieves consistent selection of the sparsity pattern and that there may be some improvement with

respect to the usual Lasso. Except for this very particular example, theoretical advantages of the group Lasso as compared to the usual Lasso were not featured in the literature. Note also that Obozinski et al. [25] focused on the consistent selection, whereas it remained unclear whether there is some improvement in the prediction properties as compared to the usual Lasso.

One of the aims of this chapter is to show that such an improvement does exist. In particular, our Theorem 5.1 implies that the prediction bound for the Group Lasso estimator that we use here is better than for the standard Lasso under the same assumptions. Furthermore, we demonstrate that as the number of tasks T increases the dependence of the bound on M disappears, provided that M grows at the rate slower than $\exp(\sqrt{T})$.

5.3 Sparsity oracle inequality

Let $1 \leq s \leq M$ be an integer that gives an upper bound on the structured sparsity $M(\beta^*)$ of the true regression vector β^*. We make the following assumption.

Assumption 5.1. *There exists a positive number $\kappa = \kappa(s)$ such that*

$$\min \left\{ \frac{\|X\Delta\|}{\sqrt{n}\|\Delta_J\|} \ : \ |J| \leq s, \Delta \in \mathbb{R}^{MT} \setminus \{0\}, \ \|\Delta_{J^c}\|_{2,1} \leq 3\|\Delta_J\|_{2,1} \right\} \geq \kappa,$$

where J^c denotes the complement of the set of indices J.

To emphasize the dependency of Assumption 5.1 on s, we will sometimes refer to it as Assumption RE(s). This is a natural extension to our setting of the Restricted Eigenvalue assumption for the usual Lasso and Dantzig selector from [6]. The ℓ_1 norms are now replaced by the mixed (2,1)-norms. Note that, however, the analogy is not complete. In fact, the sample size n in the usual Lasso setting corresponds to nT in our case, whereas in Assumption 5.1 we consider $\sqrt{\Delta^\top X^\top X\Delta/n}$ and not $\sqrt{\Delta^\top X^\top X\Delta/(nT)}$. This is done in order to have a correct normalization of κ without compulsory dependence on T (if we use the term $\sqrt{\Delta^\top X^\top X\Delta/(nT)}$ in Assumption 5.1, then $\kappa \sim T^{-1/2}$ even in the case of the identity matrix $X^\top X/n$).

Several simple sufficient conditions for Assumption 5.1 with $T = 1$ are given in [6]. Similar sufficient conditions can be stated in our more general setting. For example, it is enough to suppose that each of the matrices $X_t^\top X_t/n$ is positive definite or satisfies a Restricted Isometry condition as in [16] or the coherence condition (cf. Lemma 5.2 below).

Lemma 5.1. *Consider the model* (5.1) *for* $M \geq 2$ *and* $T, n \geq 1$. *Assume that the random vectors* $W_1, \ldots, W_T$ *are i.i.d. Gaussian with zero mean and covariance matrix* $\sigma^2 I_{n\times n}$, *all diagonal elements of the matrix* $X^\top X/n$ *are equal to* 1 *and* $M(\beta^*) \leq s$. *Let*

$$\lambda = \frac{2\sigma}{\sqrt{nT}}\left(1 + \frac{A\log M}{\sqrt{T}}\right)^{1/2},$$

where $A > 8$ *and let* $q = \min(8\log M, A\sqrt{T}/8)$. *Then with probability at least* $1 - M^{1-q}$, *for any solution* $\hat{\beta}$ *of problem* (5.3) *and all* $\beta \in \mathbb{R}^{MT}$ *we have*

$$\frac{1}{nT}\|X(\hat{\beta} - \beta^*)\|^2 + \lambda\|\hat{\beta} - \beta\|_{2,1} \leq \frac{1}{nT}\|X(\beta - \beta^*)\|^2 + 4\lambda \sum_{j\in J(\beta)} \|\hat{\beta}^j - \beta^j\|, \tag{5.6}$$

$$\frac{1}{nT}\max_{1\leq j\leq M} \|(X^\top X(\beta^* - \hat{\beta}))^j\| \leq \frac{3}{2}\lambda, \tag{5.7}$$

$$M(\hat{\beta}) \leq \frac{4\phi_{\max}}{\lambda^2 nT^2}\|X(\hat{\beta} - \beta^*)\|^2, \tag{5.8}$$

where $\phi_{\max}$ *is the maximum eigenvalue of the matrix* $X^\top X/n$.

Proof. For all $\beta \in \mathbb{R}^{MT}$, we have

$$\frac{1}{nT}\|X\hat{\beta} - y\|^2 + 2\lambda\sum_{j=1}^{M}\|\hat{\beta}^j\| \leq \frac{1}{nT}\|X\beta - y\|^2 + 2\lambda\sum_{j=1}^{M}\|\beta^j\|$$

which, using $y = X\beta^* + W$, is equivalent to

$$\frac{1}{nT}\|X(\hat{\beta} - \beta^*)\|^2 \leq \frac{1}{nT}\|X(\beta - \beta^*)\|^2 + \frac{2}{nT}W^\top X(\hat{\beta} - \beta) + 2\lambda\sum_{j=1}^{M}\left(\|\beta^j\| - \|\hat{\beta}^j\|\right). \tag{5.9}$$

By Hölder's inequality, we have that

$$W^\top X(\hat{\beta} - \beta) \leq \|X^\top W\|_{2,\infty}\|\hat{\beta} - \beta\|_{2,1}$$

where

$$\|X^\top W\|_{2,\infty} = \max_{1\leq j\leq M}\sqrt{\sum_{t=1}^{T}\left(\sum_{i=1}^{n}(x_{ti})_j W_{ti}\right)^2}.$$

Consider the random event

$$\mathcal{A} = \left\{\frac{1}{nT}\|X^\top W\|_{2,\infty} \leq \frac{\lambda}{2}\right\}.$$

Since we assume all diagonal elements of the matrix $X^\top X/n$ to be equal to 1, the random variables

$$V_{tj} = \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(x_{ti})_j W_{ti},$$

$t = 1, \ldots, T$, are *i.i.d.* standard Gaussian. Using this fact we can write, for any $j = 1, \ldots, M$,

$$\begin{aligned} \Pr\left(\sum_{t=1}^{T}\left(\sum_{i=1}^{n}(x_{ti})_j W_{ti}\right)^2 \geq \frac{\lambda^2(nT)^2}{4}\right) &= \Pr\left(\chi_T^2 \geq \frac{\lambda^2 nT^2}{4\sigma^2}\right) \\ &= \Pr\left(\chi_T^2 \geq T + A\sqrt{T}\log M\right), \end{aligned}$$

where χ_T^2 is a chi-square random variable with T degrees of freedom. We now apply Lemma 5.4, the union bound and the fact that $A > 8$ to get

$$\Pr(\mathcal{A}^c) \leq M \exp\left(-\frac{A\log M}{8}\min\left(\sqrt{T}, A\log M\right)\right) \leq M^{1-q}.$$

It follows from (5.9) that, on the event $\mathcal{A}$.

$$\begin{aligned} &\frac{1}{nT}\|X(\hat{\beta} - \beta^*)\|^2 + \lambda\sum_{j=1}^{M}\|\hat{\beta}^j - \beta^j\| \leq \\ &\frac{1}{nT}\|X(\beta - \beta^*)\|^2 + 2\lambda\sum_{j=1}^{M}\left(\|\hat{\beta}^j - \beta^j\| + \|\beta^j\| - \|\hat{\beta}^j\|\right) \\ &\leq \frac{1}{nT}\|X(\beta - \beta^*)\|^2 + 4\lambda\sum_{j\in J(\beta)}\|\hat{\beta}^j - \beta^j\|, \end{aligned}$$

which coincides with (5.6). To prove (5.7), we use the inequality

$$\frac{1}{nT}\max_{1\leq j\leq M}\|(X^\top(y - X\hat{\beta}))^j\| \leq \lambda, \tag{5.10}$$

which follows from (5.4) and (5.5). Then,

$$\frac{1}{nT}\|(X^\top(X(\hat{\beta} - \beta^*)))^j\| \leq \frac{1}{nT}\|(X^\top(X\hat{\beta} - y))^j\| + \frac{1}{nT}\|(X^\top W)^j\|,$$

where we have used $y = X\beta^* + W$ and the triangle inequality. The result then follows by combining the last inequality with inequality (5.10) and using the definition of the event $\mathcal{A}$.

Finally, we prove (5.8). First, observe that, on the event $\mathcal{A}$,

$$\frac{1}{nT}\|(X^\top X(\hat{\beta} - \beta^*))^j\| \geq \frac{\lambda}{2}, \quad \text{if } \hat{\beta}^j \neq 0.$$

This fact follows from (5.4), (5.2) and the definition of the event $\mathcal{A}$. The following chain

yields the result:

$$\begin{aligned} M(\hat{\beta}) &\leq \frac{4}{\lambda^2(nT)^2} \sum_{j \in J(\hat{\beta})} \|(X^\top X(\hat{\beta} - \beta^*))^j\|^2 \\ &\leq \frac{4}{\lambda^2(nT)^2} \sum_{j=1}^{M} \|(X^\top X(\hat{\beta} - \beta^*))^j\|^2 \\ &= \frac{4}{\lambda^2(nT)^2} \|X^\top X(\hat{\beta} - \beta^*)\|^2 \\ &\leq \frac{4\phi_{\max}}{\lambda^2 nT^2} \|X(\hat{\beta} - \beta^*)\|^2, \end{aligned}$$

where, in the last line we have used the fact that the eigenvalues of $X^\top X/n$ are bounded from above by $\Phi_{\max}$. □

We are now ready to state the main result of this section.

Theorem 5.1. *Consider the model* (5.1) *for* $M \geq 2$ *and* $T, n \geq 1$. *Assume that the random vectors* $W_1, \ldots, W_T$ *are i.i.d. Gaussian with zero mean and covariance matrix* $\sigma^2 I_{n \times n}$, *all diagonal elements of the matrix* $X^\top X/n$ *are equal to* 1 *and* $M(\beta^*) \leq s$. *Furthermore let Assumption 5.1 hold with* $\kappa = \kappa(s)$ *and let* $\phi_{\max}$ *be the largest eigenvalue of the matrix* $X^\top X/n$. *Let*

$$\lambda = \frac{2\sigma}{\sqrt{nT}} \left(1 + \frac{A \log M}{\sqrt{T}}\right)^{1/2},$$

where $A > 8$ *and let* $q = \min(8 \log M, A\sqrt{T}/8)$. *Then with probability at least* $1 - M^{1-q}$, *for any solution* $\hat{\beta}$ *of problem* (5.3) *we have*

$$\frac{1}{nT} \|X(\hat{\beta} - \beta^*)\|^2 \leq \frac{64\sigma^2}{\kappa^2} \frac{s}{n} \left(1 + \frac{A \log M}{\sqrt{T}}\right) \tag{5.11}$$

$$\frac{1}{\sqrt{T}} \|\hat{\beta} - \beta^*\|_{2,1} \leq \frac{32\sigma}{\kappa^2} \frac{s}{\sqrt{n}} \sqrt{1 + \frac{A \log M}{\sqrt{T}}} \tag{5.12}$$

$$M(\hat{\beta}) \leq \frac{64\phi_{\max}}{\kappa^2} s. \tag{5.13}$$

If, in addition, Assumption RE(2s) holds, then with the same probability for any solution $\hat{\beta}$ *of problem* (5.3) *we have*

$$\frac{1}{\sqrt{T}} \|\hat{\beta} - \beta^*\| \leq \frac{8\sqrt{10}\sigma}{\kappa^2(2s)} \sqrt{\frac{s}{n}} \sqrt{1 + \frac{A \log M}{\sqrt{T}}}. \tag{5.14}$$

Proof. We act similarly to the proof of Theorem 6.2 in [6]. Let $J = J(\beta^*) = \{j : (\beta^*)^j \neq 0\}$. By inequality (5.6) with $\beta = \beta^*$ we have, on the event $\mathcal{A}$, that

$$\begin{aligned} \frac{1}{nT}\|X(\hat{\beta} - \beta^*)\|^2 &\leq 4\lambda \sum_{j \in J} \|\hat{\beta}^j - \beta^{*j}\| \\ &\leq 4\lambda\sqrt{s}\|(\hat{\beta} - \beta^*)_J\|. \end{aligned} \tag{5.15}$$

Moreover by the same inequality, on the event $\mathcal{A}$, we have $\sum_{j=1}^M \|\hat{\beta}^j - \beta^{*j}\| \leq 4\sum_{j \in J} \|\hat{\beta}^j - \beta^{*j}\|$, which implies that $\sum_{j \in J^c} \|\hat{\beta}^j - \beta^{*j}\| \leq 3\sum_{j \in J} \|\hat{\beta}^j - \beta^{*j}\|$. Thus, by Assumption 5.1

$$\|(\hat{\beta} - \beta^*)_J\| \leq \frac{\|X(\hat{\beta} - \beta^*)\|}{\kappa\sqrt{n}}. \tag{5.16}$$

Now, (5.11) follows from (5.15) and (5.16). Inequality (5.12) follows again by noting that

$$\sum_{j=1}^M \|\hat{\beta}^j - \beta^{*j}\| \leq 4\sum_{j \in J} \|\hat{\beta}^j - \beta^{*j}\| \leq 4\sqrt{s}\|(\hat{\beta} - \beta^*)_J\|$$

and then using (5.11). Inequality (5.13) follows from (5.8) and (5.11).

Finally, we prove (5.14). Let $\Delta = \hat{\beta} - \beta^*$ and let J' be the set of indices in J^c corresponding to s maximal in absolute value norms $\|\Delta^j\|$. Consider the set $J_{2s} = J \cup J'$. Note that $|J_{2s}| \leqslant 2s$. Let $\|\Delta_{J^c}^{(k)}\|$ denote the k-th largest norm in the set $\{\|\Delta^j\| : j \in J^c\}$. Then, clearly,

$$\|\Delta_{J^c}^{(k)}\| \leq \sum_{j \in J^c} \|\Delta^j\|/k = \|\Delta_{J^c}\|_{2,1}/k.$$

This and the fact that $\|\Delta_{J^c}\|_{2,1} \leq 3\|\Delta_J\|_{2,1}$ on the event $\mathcal{A}$ implies

$$\begin{aligned} \sum_{j \in J_{2s}^c} \|\Delta^j\|^2 &\leq \sum_{k=s+1}^{\infty} \frac{\|\Delta_{J^c}\|_{2,1}^2}{k^2} \\ &\leq \frac{\|\Delta_{J^c}\|_{2,1}^2}{s} \leq \frac{9\|\Delta_J\|_{2,1}^2}{s} \\ &\leq 9\sum_{j \in J} \|\Delta^j\|^2 \leq 9\sum_{j \in J_{2s}} \|\Delta^j\|^2. \end{aligned}$$

Therefore, on $\mathcal{A}$ we have

$$\|\Delta\|^2 \leq 10\sum_{j \in J_{2s}} \|\Delta^j\|^2 \equiv 10\|\Delta_{J_{2s}}\|^2 \tag{5.17}$$

and also from (5.15):

$$\frac{1}{nT}\|X\Delta\|^2 \leq 4\lambda\sqrt{s}\|\Delta_{J_{2s}}\|. \tag{5.18}$$

In addition, $\|\Delta_{J^c}\|_{2,1} \leq 3\|\Delta_J\|_{2,1}$ easily implies that

$$\|\Delta_{J_{2s}^c}\|_{2,1} \leq 3\|\Delta_{J_{2s}}\|_{2,1}.$$

Combining (5.18) with Assumption RE($2s$) we find that on the event $\mathcal{A}$ it holds that

$$\|\Delta_{J_{2s}}\| \leq \frac{4\lambda\sqrt{s}\,T}{\kappa^2(2s)}\,.$$

This inequality and (5.17) yield (5.14). □

Theorem 5.1 is valid for any fixed n, M, T; the approach is non-asymptotic. Some relations between these parameters are relevant in the particular applications and various asymptotics can be derived as corollaries. For example, in multi-task learning it is natural to assume that $T \geq n$, and the motivation for our approach is the strongest if also $M \gg n$. The bounds of Theorem 5.1 are meaningful if the sparsity index s is small as compared to the sample size n and the logarithm of the dimension $\log M$ is not too large as compared to $\sqrt{T}$.

Note also that the values T and $\sqrt{T}$ in the denominators of the right-hand sides of (5.11), (5.12), and (5.14) appear quite naturally. For instance, the norm $\|\hat{\beta} - \beta^*\|_{2,1}$ in (5.12) is a sum of M terms each of which is a Euclidean norm of a vector in $\mathbb{R}^T$, and thus it is of the order $\sqrt{T}$ if all the components are equal. Therefore, (5.12) can be interpreted as a correctly normalized “error per coefficient" bound.

We now state several important conclusions. They are all valid for the general Group Lasso, and not only in the multi-task learning setup. They key point for their validity is the structured sparsity assumption.

1. *Theorem 5.1 applies to the general Group Lasso setting.* Indeed, the proofs in this section do not use the fact that the matrix $X^\top X$ is block-diagonal. The only restriction on $X^\top X$ is given in Assumption 3.1. For example, Assumption 3.1 is obviously satisfied if $X^\top X/(nT)$ (the correctly normalized Gram matrix of the regression model (5.2)) has a positive minimal eigenvalue.

2. *The dependence on the dimension M is negligible for large T.* Indeed, the bounds of Theorem 5.1 become independent of M if we choose the number of tasks T larger than $\log^2 M$. A striking fact is that no relation between the sample size n and the dimension M is required. This is quite in contrast to the previous results on sparse recovery where the assumption $\log M = o(n)$ was considered as *sine qua non* constraint. For example, Theorem 5.1 gives meaningful bounds if $M = \exp(n^\gamma)$ for arbitrarily large $\gamma > 0$,

provided that $T > n^{2\gamma}$. This is due to the structured sparsity assumption, and is not conditioned by the block-diagonal (multi-task) structure of the regression matrices.

3. *Our estimator admits better risk bounds than the usual Lasso.* Let us explain this point considering the example of the prediction error bound (5.11). Indeed, for the same multi-task setup, we can apply a usual Lasso estimator $\hat{\beta}^L$, that is a solution of the following optimization problem

$$\min_{\beta} S(\beta) + 2\lambda' \sum_{t=1}^{T} \sum_{j=1}^{M} |\beta_{tj}|$$

where $\lambda' > 0$ is a tuning parameter. We will use the bounds of [6] for the prediction error of $\hat{\beta}^L$. For a fair comparison with Theorem 5.1, we assume that we are in the most favorable situation where $M < n$, each of the matrices $\frac{1}{n} X_t^T X_t$ is positive definite and has minimal eigenvalue greater than κ^2. This implies both Assumption 3.1 and the Restricted Eigenvalue assumption as stated in [6]. Next, we assume, as in Theorem 5.1, that the diagonal elements of the matrix $X^\top X/n$ are equal to 1.

To use the results of [6], we note that the parameters n, M, s therein correspond to $n' = nT$, $M' = MT$, $s' = sT$ in our setup, and the minimal eigenvalue of the matrix $\frac{1}{n'} X^T X = \frac{1}{nT} X^T X$ is greater than $(\kappa')^2 \equiv \kappa^2/T$. Another particularity is that, due to our normalization, the diagonal elements of the matrix $\frac{1}{nT} X^T X$ are equal to $1/T$, and not to 1, as in [6]. This results in the fact that the correct λ' is by a $\sqrt{T}$ factor smaller than that given in [6]:

$$\lambda' = A' \frac{\sigma}{\sqrt{T}} \sqrt{\frac{\log(MT)}{nT}},$$

where $A' > 2\sqrt{2}$. We can then act as in the proof of inequality (7.8) from [6] (cf. (B.31) in [6]) to obtain that, with probability at least $1 - (MT)^{1 - \frac{(A')^2}{8}}$, it holds

$$\begin{aligned} \frac{1}{nT} ||X(\hat{\beta}^L - \beta^*)||^2 &\leqslant \frac{16 s' (\lambda')^2}{(\kappa')^2} \\ &= \frac{16 (A')^2}{\kappa^2} \sigma^2 s \frac{\log(MT)}{n}. \end{aligned} \tag{5.19}$$

Comparing with (5.11) we conclude that if $\log M$ is not too large as compared to $\sqrt{T}$ the rate of prediction bound (5.11) for the Group Lasso is by a factor of $\log(MT)$ better than for the usual Lasso under the same assumptions. Let us emphasize that the improvement is only due to the property that β^* is structured sparse.

Finally, we note that [83] follow the scheme of the proof of [6] to derive similar in spirit to ours but coarse oracle inequalities. Their results do not explain the advantages discussed in

the points 1–3 above. Indeed, the tuning parameter λ chosen in [83], pp. 614–615, is larger than our λ by at least a factor of $\sqrt{T}$. As a consequence, the corresponding bounds in the oracle inequalities of [83] are larger than ours by positive powers of T.

5.4 Coordinate-wise estimation and selection of sparsity pattern

In this section, we show how from any solution of the problem (5.3) we can reliably estimate the correct sparsity pattern with high probability.

We first introduce some more notation. We define the Gram matrix of the design $\Psi = \frac{1}{n}X^\top X$. Note that Ψ is a $MT \times MT$ block-diagonal matrix with T blocks of dimension $M \times M$ each. We denote these blocks by $\Psi_t = \frac{1}{n}X_t^\top X_t \equiv (\Psi_{tj,tk})_{j,k=1,\dots,M}$.

In this section we assume that the following condition holds true.

Assumption 5.2. *The elements $\Psi_{tj,tk}$ of the Gram matrix Ψ satisfy*

$$\Psi_{tj,tj} = 1, \ \ \forall 1 \leqslant j \leqslant M,\, 1 \leqslant t \leqslant T,$$

and

$$\max_{1 \leqslant t \leqslant T, j \neq k} |\Psi_{tj,tk}| \leqslant \frac{1}{7\alpha s},$$

for some integer $s \geqslant 1$ and some constant $\alpha > 1$.

Note that the above assumption on Ψ implies Assumption 5.1 as we prove in the following lemma.

Lemma 5.2. *Let Assumption 5.2 be satisfied. Then Assumption 5.1 is satisfied with $\kappa = \sqrt{1 - \frac{1}{\alpha}}$.*

Proof. For any subset J of $\{1, \dots, M\}$ such that $|J| \leqslant s$ and any $\Delta \in \mathbb{R}^{MT}$ such that $\|\Delta_{J^c}\|_{2,1} \leqslant 3\|\Delta_J\|_{2,1}$, we have

$$\begin{aligned}
\frac{\Delta_J^\top \Psi \Delta_J}{\|\Delta_J\|^2} &= 1 + \frac{\Delta_J^\top (\Psi - I_{MT \times MT})\Delta_J}{\|\Delta_J\|^2} \\
&\geqslant 1 - \frac{1}{7\alpha s} \frac{\Big(\sum_{j \in J} \sum_{t=1}^T |\Delta_{tj}|\Big)^2}{\|\Delta_J\|^2} \\
&\geqslant 1 - \frac{1}{7\alpha}
\end{aligned}$$

where we have used Assumption 5.2 and the Cauchy-Schwarz inequality. Next, using consecutively Assumption 5.2, the Cauchy-Schwarz inequality and the inequality $\|\Delta_{J^c}\|_{2,1} \leqslant 3\|\Delta_J\|_{2,1}$ we obtain

$$\begin{aligned}
\frac{|\Delta_{J^c}^\top \Psi \Delta_J|}{\|\Delta_J\|^2} &\leqslant \frac{1}{7\alpha s} \frac{\sum_{t=1}^T \sum_{j\in J} \sum_{k\in J^c} |\Delta_{tj}||\Delta_{tk}|}{\|\Delta_J\|^2} \\
&\leqslant \frac{1}{7\alpha s} \frac{\sum_{j\in J, k\in J^c} \|\Delta^j\| \|\Delta^k\|}{\|\Delta_J\|^2} \\
&\leqslant \frac{3}{7\alpha s} \frac{\|\Delta_J\|_{2,1}^2}{\|\Delta_J\|^2} \\
&\leqslant \frac{3}{7\alpha}.
\end{aligned}$$

Combining these inequalities we find

$$\frac{\Delta^\top \Psi \Delta}{\|\Delta_J\|^2} \geqslant \frac{\Delta_J^\top \Psi \Delta_J}{\|\Delta_J\|^2} + \frac{2\Delta_{J^c}^\top \Psi \Delta_J}{\|\Delta_J\|^2} \geqslant 1 - \frac{1}{\alpha} > 0.$$

□

Note also that, by an argument as in [71], it is not hard to show that under Assumption 5.2 the vector β^* satisfying (5.2) is unique.

Theorem 5.1 provides bounds for compound measures of risk, that is, depending simultaneously on all the vectors β^j. An important question is to evaluate the performance of estimators for each of the components β^j separately. The next theorem provides a bound of this type and, as a consequence, a result on the selection of sparsity pattern.

Theorem 5.2. *Consider the model (5.1) for $M \geqslant 2$ and $T, n \geqslant 1$. Let the assumptions of Lemma 5.1 be satisfied and let Assumption 5.2 hold with the same s. Set*

$$c = \left(3 + \frac{32}{7(\alpha - 1)}\right)\sigma.$$

Let λ, A and $W_1, \ldots, W_T$ be as in Lemma 5.1. Then with probability at least $1 - M^{1-q}$, where $q = \min(8 \log M, ByA\sqrt{T}/8)$, for any solution $\hat{\beta}$ of problem (5.3) we have

$$\frac{1}{\sqrt{T}} \|\hat{\beta} - \beta^*\|_{2,\infty} \leqslant \frac{c}{\sqrt{n}} \sqrt{1 + \frac{A \log M}{\sqrt{T}}}. \tag{5.20}$$

If, in addition,

$$\min_{j \in J(\beta^*)} \frac{1}{\sqrt{T}} \|(\beta^*)^j\| > \frac{2c}{\sqrt{n}} \sqrt{1 + \frac{A \log M}{\sqrt{T}}}, \tag{5.21}$$

then with the same probability for any solution $\hat{\beta}$ of problem (5.3) the set of indices

$$\hat{J} = \left\{ j : \frac{1}{\sqrt{T}} \|\hat{\beta}^j\| > \frac{c}{\sqrt{n}} \sqrt{1 + \frac{A \log M}{\sqrt{T}}} \right\} \tag{5.22}$$

estimates correctly the sparsity pattern $J(\beta^*)$, that is,

$$\hat{J} = J(\beta^*).$$

Proof. Set $\Delta = \hat{\beta} - \beta^*$. We have

$$\|\Delta\|_{2,\infty} \leqslant \|\Psi\Delta\|_{2,\infty} + \|(\Psi - I_{MT\times MT})\Delta\|_{2,\infty}. \tag{5.23}$$

Using Assumption 5.2 we obtain

$$\begin{aligned}
\|(\Psi - I_{MT\times MT})\Delta\|_{2,\infty} &= \max_{1\leqslant j\leqslant M} \left[\sum_{t=1}^{T} \left(\sum_{k=1:k\neq j}^{M} |\Psi_{tj,tk}||\Delta_{tk}| \right)^2 \right]^{1/2} \\
&\leqslant \max_{1\leqslant j\leqslant M} \left[\max_{1\leqslant t\leqslant T, j\neq k} |\Psi_{tj,tk}|^2 \sum_{t=1}^{T} \left(\sum_{k=1:k\neq j}^{M} |\Delta_{tk}| \right)^2 \right]^{1/2} \\
&\leqslant \frac{1}{7\alpha s} \left[\sum_{t=1}^{T} \left(\sum_{k=1}^{M} |\Delta_{tk}| \right)^2 \right]^{1/2}.
\end{aligned} \tag{5.24}$$

By the Minkowski inequality for the Euclidean norm in $\mathbb{R}^T$,

$$\left[\sum_{t=1}^{T} \left(\sum_{k=1}^{M} |\Delta_{tk}| \right)^2 \right]^{1/2} \leqslant \|\Delta\|_{2,1}. \tag{5.25}$$

Combining the three above displays we get

$$\|\Delta\|_{2,\infty} \leqslant \|\Psi\Delta\|_{2,\infty} + \frac{1}{7\alpha s} \|\Delta\|_{2,1}.$$

Thus, by Lemma 5.1 and Theorem 5.1, with probability at least $1 - M^{1-q}$,

$$\|\Delta\|_{2,\infty} \leqslant \left(\frac{3}{2} + \frac{16}{7\alpha\kappa^2} \right) \lambda T.$$

By Lemma 5.2, $\alpha\kappa^2 = \alpha - 1$, which yields the first result of the theorem. The second result follows from the first one in an obvious way. □

Assumption of type (5.21) is inevitable in the context of selection of sparsity pattern. It says that the vectors $(\beta^*)^j$ cannot be arbitrarily close to 0 for j in the pattern. Their norms should be at least somewhat larger than the noise level.

The second result of Theorem 5.2 (selection of sparsity pattern) can be compared with [4, 83] who considered the Group Lasso. There are several differences. First, our estimator $\hat{J}$ is based on thresholding of the norms $\|\hat{\beta}^j\|$, while [4, 83] take instead the set where these norms do not vanish. In practice, the latter is known to be a poor selector; it typically overestimates the true sparsity pattern. Second, [4, 83] consider specific asymptotic settings, while our result holds for any fixed n, M, T. Different kinds of asymptotics can be therefore obtained as simple corollaries. Finally, note that the estimator $\hat{\beta}$ is not necessarily unique. Though [83] does not discuss this fact, the proof there only shows that *there exists a subsequence of solutions $\hat{\beta}$ of (5.3)* such that the set $\{j : \|\hat{\beta}^j\| \neq 0\}$ coincides with the sparsity pattern $J(\beta^*)$ in some specified asymptotics (we note that the "if and only if" claim before formula (23) in [83] is not proved). In contrast, the argument in Theorem 5.2 does not require any analysis of the uniqueness issues, though it is not excluded that the solution is indeed unique. It guarantees that *simultaneously for all solutions $\hat{\beta}$ of (5.3)* and any fixed n, M, T the correct selection is done with high probability.

Theorems 5.1 and 5.2 imply the following corollary.

Corollary 5.1. *Consider the model (5.1) for $M \geqslant 2$ and $T, n \geqslant 1$. Let the assumptions of Lemma 5.1 be satisfied and let Assumption 5.2 holds with the same s. Let λ, A and $W_1, \dots, W_T$ be as in Lemma 5.1. Then with probability at least $1 - M^{1-q}$, where $q = \min(8 \log M, A\sqrt{T}/8)$, for any solution $\hat{\beta}$ of problem (5.3) and any $1 \leq p < \infty$ we have*

$$\frac{1}{\sqrt{T}} \|\hat{\beta} - \beta^*\|_{2,p} \leqslant c_1 \sigma \frac{s^{1/p}}{\sqrt{n}} \sqrt{1 + \frac{A \log M}{\sqrt{T}}}\,, \tag{5.26}$$

where

$$c_1 = \left(\frac{32\alpha}{\alpha - 1}\right)^{1/p} \left(3 + \frac{32}{7(\alpha - 1)}\right)^{1 - \frac{1}{p}}.$$

If, in addition, (5.21) holds, then with the same probability for any solution $\hat{\beta}$ of problem (5.3) and any $1 \leq p < \infty$ we have

$$\frac{1}{\sqrt{T}} \|\hat{\beta} - \beta^*\|_{2,p} \leqslant c_1 \sigma \frac{|\hat{J}|^{1/p}}{\sqrt{n}} \sqrt{1 + \frac{A \log M}{\sqrt{T}}}\,, \tag{5.27}$$

where $\hat{J}$ is defined in (5.22).

Proof. Set $\Delta = \hat{\beta} - \beta$. For any $p \geqslant 1$ we have

$$\frac{1}{\sqrt{T}}\|\Delta\|_{2,p} \leqslant \left(\frac{1}{\sqrt{T}}\|\Delta\|_{2,1}\right)^{\frac{1}{p}} \left(\frac{1}{\sqrt{T}}\|\Delta\|_{2,\infty}\right)^{1-\frac{1}{p}}.$$

Combining (5.12), (5.20) with $\kappa = \sqrt{1-\frac{1}{\alpha}}$ and the above display yields the first result. □

Inequalities (5.20) and (5.27) provide confidence intervals for the unknown parameter β^* in mixed (2,p)-norms.

For averages of the coefficients β_{tj} we can establish a sign consistency result which is somewhat stronger than the result in Theorem 5.2. For any $\beta \in \mathbb{R}^M$, define $\overrightarrow{\operatorname{sign}}(\beta) = (\operatorname{sign}(\beta^1), \ldots, \operatorname{sign}(\beta^M))^\top$ where

$$\operatorname{sign}(t) = \begin{cases} 1 & \text{if } t > 0, \\ 0 & \text{if } t = 0, \\ -1 & \text{if } t < 0. \end{cases}$$

Introduce the averages

$$a_j^* = \frac{1}{T}\sum_{t=1}^{T} \beta_{tj}^*, \quad \hat{a}_j = \frac{1}{T}\sum_{t=1}^{T} \hat{\beta}_{tj}.$$

Consider the threshold $\tau = \frac{c}{\sqrt{n}}\sqrt{1 + \frac{A\log M}{\sqrt{T}}}$ and define a thresholded estimator

$$\tilde{a}_j = \hat{a}_j I\big\{|\hat{a}_j| > \tau\big\}.$$

Let $\tilde{a}$ and a^* be the vectors with components $\tilde{a}_j$ and a_j^*, $j = 1, \ldots, M$, respectively. We need the following additional assumption.

Assumption 5.3. *It holds that*

$$\min_{j \in J(a^*)} |a_j^*| \geq \frac{2c}{\sqrt{n}}\sqrt{1 + \frac{A\log M}{\sqrt{T}}}.$$

This assumption says that we cannot recover arbitrarily small components. Similar assumptions are standard in the literature on sign consistency (see, for example, [71] for more details and references).

Theorem 5.3. *Consider the model (5.1) for $M \geqslant 2$ and $T, n \geqslant 1$. Let the assumptions of Lemma 5.1 be satisfied and let Assumption 5.2 hold with the same s. Let λ and A be defined as in Lemma 5.1 and c as in Theorem 5.2. Then with probability at least $1 - M^{1-q}$, where $q = \min(8\log M, A\sqrt{T}/8)$, for any solution $\hat{\beta}$ of problem (5.3) we have*

$$\max_{1 \leqslant j \leqslant M} |\hat{a}_j - a_j^*| \leqslant \frac{c}{\sqrt{n}}\sqrt{1 + \frac{A\log M}{\sqrt{T}}}.$$

If, in addition, Assumption 5.3 holds, then with the same probability, for any solution $\hat{\beta}$ of problem (5.3), $\tilde{a}$ recovers the sign pattern of a^*:

$$\overrightarrow{\operatorname{sign}}(\tilde{a}) = \overrightarrow{\operatorname{sign}}(a^*).$$

Proof. Note that for every $j \in \mathbb{N}_M$

$$|\hat{a}_j - a_j^*| \leq \frac{1}{\sqrt{T}} \|\hat{\beta} - \beta^*\|_{2,\infty} \leq \frac{c}{\sqrt{n}} \sqrt{1 + \frac{A \log M}{\sqrt{T}}}.$$

The proof is then similar to that of Theorem 5.2. □

5.5 Non-Gaussian noise

In this section, we only assume that the random variables $W_{ti}, i \in \mathbb{N}_n, t \in \mathbb{N}_T$, are independent with zero mean and finite fourth moment $\mathbb{E}[W_{ti}^4] \leqslant \sigma^4$. In this case the results remain similar to those of the previous sections, though the concentration effect is weaker. We need the following technical assumption

Assumption 5.4. *The matrix X is such that*

$$\max_{t \in \mathbb{N}_T} \left(\frac{1}{n} \sum_{i=1}^{n} \max_{j \in \mathbb{N}_M} |(x_{ti})_j|^4 \right) \leqslant c'^4,$$

for a constant $c' > 0$.

This assumption is quite mild. It is satisfied for example, if all $(x_{ti})_j$ are bounded in absolute value by a constant uniformly in i, t, j. We have the two following theorems.

Theorem 5.4. *Consider the model (5.1) for $M \geq 3$ and $T, n \geq 1$. Assume that the random vectors $W_1, \ldots, W_T$ are independent with zero mean and finite fourth moment $\mathbb{E}[W_{ti}^4] \leqslant \sigma^4$, all diagonal elements of the matrix $X^\top X/n$ are equal to 1 and $M(\beta^*) \leq s$. Let also Assumption 5.4 be satisfied. Furthermore let κ be defined as in Assumption 5.1 and $\phi_{\max}$ be the largest eigenvalue of the matrix $X^\top X/n$. Let*

$$\lambda = \frac{c'\sigma}{\sqrt{nT}} \left(1 + \frac{(\log M)^{3+\delta}}{\sqrt{T}} \right)^{1/2},$$

with $\delta > 0$. Then with probability at least $1 - \frac{(2e\log M - e)(8\log(14M))^2}{(\log M)^{3+\delta}}$, for any solution $\hat{\beta}$ of problem (5.3) we have

$$\frac{1}{nT}\|X(\hat{\beta} - \beta^*)\|^2 \leq \frac{16}{\kappa^2} c'^2 \sigma^2 \frac{s}{n} \left(1 + \frac{(\log M)^{3+\delta}}{\sqrt{T}}\right), \tag{5.28}$$

$$\frac{1}{\sqrt{T}}\|\hat{\beta} - \beta^*\|_{2,1} \leq \frac{16}{\kappa^2} c'\sigma \frac{s}{\sqrt{n}} \left(1 + \frac{(\log M)^{3+\delta}}{\sqrt{T}}\right)^{1/2}, \tag{5.29}$$

$$M(\hat{\beta}) \leq \frac{64\phi_{\max}}{\kappa^2} s. \tag{5.30}$$

If, in addition, Assumption RE(2s) holds, then with the same probability for any solution $\hat{\beta}$ of problem (5.3) we have

$$\frac{1}{\sqrt{T}}\|\hat{\beta} - \beta^*\| \leq \frac{4\sqrt{10}c'\sigma}{\kappa^2(2s)} \sqrt{\frac{s}{n}} \left(1 + \frac{(\log M)^{3+\delta}}{\sqrt{T}}\right)^{1/2}.$$

Theorem 5.5. *Consider the model (5.1) for $M \geqslant 3$ and $T, n \geqslant 1$. Let the assumptions of Theorem 5.4 be satisfied and let Assumption 5.2 hold with the same s. Set*

$$c = \left(\frac{3}{2} + \frac{16}{7(\alpha - 1)}\right) c'\sigma.$$

Let λ be as in Theorem 5.4. Then with probability at least $1 - \frac{(2e\log M - e)(8\log(14M))^2}{(\log M)^{3+\delta}}$, for any solution $\hat{\beta}$ of problem (5.3) we have

$$\frac{1}{\sqrt{T}}\|\hat{\beta} - \beta^*\|_{2,\infty} \leqslant c\frac{1}{\sqrt{n}} \left(1 + \frac{(\log M)^{3+\delta}}{\sqrt{T}}\right)^{1/2}.$$

If, in addition, it holds that

$$\min_{j \in J(\beta^*)} \frac{1}{\sqrt{T}}\|(\beta^*)^j\| > 2c\frac{1}{\sqrt{n}} \left(1 + \frac{(\log M)^{3+\delta}}{\sqrt{T}}\right)^{1/2},$$

then with the same probability for any solution $\hat{\beta}$ of problem (5.3) the set of indices

$$\hat{J} = \left\{j : \frac{1}{\sqrt{T}}\|\hat{\beta}^j\| > c\frac{1}{\sqrt{n}} \left(1 + \frac{(\log M)^{3+\delta}}{\sqrt{T}}\right)^{1/2}\right\}$$

estimates correctly the sparsity pattern $J(\beta^*)$:

$$\hat{J} = J(\beta^*).$$

Proof. The proofs of these theorems are similar to those of Theorems 5.1 and 5.2 up to a modification of the bound on $P(\mathcal{A}^c)$ in Lemma 5.1. We consider now the event

$$\mathcal{A} = \left\{\max_{j=1}^{M} \sqrt{\sum_{t=1}^{T} \left(\sum_{i=1}^{n} (x_{ti})_j W_{ti}\right)^2} \leq \lambda nT\right\}.$$

Define the random variables

$$Y_{tj} = \left(\sum_{i=1}^{n} (x_{ti})_j W_{ti} \right)^2 - \sum_{i=1}^{n} |(x_{ti})_j|^2 \mathbb{E}[W_{ti}^2], \ \forall j, t.$$

We have

$$\begin{aligned} P(\mathcal{A}^c) &= P\left(\max_{1 \leqslant j \leqslant M} \sum_{t=1}^{T} \left(\sum_{i=1}^{n} (x_{ti})_j W_{ti} \right)^2 \geqslant (\lambda n T)^2 \right) \\ &\leqslant P\left(\max_{1 \leqslant j \leqslant M} \sum_{t=1}^{T} Y_{tj} \geqslant c'^2 \sigma^2 n \sqrt{T} (\log M)^{3+\delta} \right) \\ &\leqslant \frac{\mathbb{E}\left(\max_{1 \leqslant j \leqslant M} \left| \sum_{t=1}^{T} Y_{tj} \right|^2 \right)}{c'^4 \sigma^4 n^2 T (\log M)^{3+\delta}}, \end{aligned}$$

where we have used the Markov inequality in the last line. Lemma 5.5 below yields

$$\mathbb{E}\left(\max_{1 \leqslant j \leqslant M} \left| \sum_{t=1}^{T} Y_{tj} \right|^2 \right) \leqslant (2e \log M - e) \sum_{t=1}^{T} \mathbb{E}\left(\max_{1 \leqslant j \leqslant M} |Y_{tj}|^2 \right).$$

Applying Lemma 5.3 below with the constant $c(4) = 7$ when $M \geqslant 3$ yields

$$\begin{aligned} \mathbb{E}\left(\max_{1 \leqslant j \leqslant M} |Y_{tj}|^2 \right) &\leqslant \mathbb{E}\left(\max_{1 \leqslant j \leqslant M} \left| \sum_{i=1}^{n} (x_{ti})_j W_{ti} \right|^4 \right) \\ &\leqslant (8 \log(14M))^2 \mathbb{E}\left(\sum_{i=1}^{n} \max_{1 \leqslant j \leqslant M} (x_{ti})_j^2 W_{ti}^2 \right)^2 \\ &\leqslant (8 \log(14M))^2 n \sum_{i=1}^{n} \max_{1 \leqslant j \leqslant M} (x_{ti})_j^4 \sigma^4 \end{aligned}$$

Combining the above three displays yields

$$P(\mathcal{A}^c) \leqslant \frac{(2e \log M - e)(8 \log(14M))^2}{(\log M)^{3+\delta}}.$$

□

5.6 Nemirovski moment inequality

In this section, we prove an inequality for the m-th moment of maxima of sums of independent random variables. The case $m = 2$ is - modulo constants - Nemirovski's inequality

(see [36], Corollary 2.4 page 5). The latter actually concerns the second moment of ℓ_p-norms ($1 \leqslant p \leqslant \infty$) of sums of independent random variables in $\mathbb{R}^M$, whereas we only consider the case $p = \infty$.

Lemma 5.3. ("Nemirovski moment inequality") *Let $Z_1, \ldots, Z_n$ be independent vectors in $\mathbb{R}^M$. Then for $m \geqslant 1$, there exists a constant $c(m) \geqslant 2$ such that for any $M \geqslant 1$, we have*

$$\mathbb{E} \max_{1 \leqslant j \leqslant M} \left| \sum_{i=1}^{n} \Big(Z_{i,j} - \mathbb{E} Z_{i,j} \Big) \right|^m \leq \Big[8 \log(2c(m)M) \Big]^{m/2} \mathbb{E} \Big[\sum_{i=1}^{n} \max_{1 \leqslant j \leqslant M} Z_{i,j}^2 \Big]^{m/2}.$$

Proof. Let $(\varepsilon_1, \ldots, \varepsilon_n)$ be a Rademacher sequence independent of $\mathbf{Z} := (Z_1, \ldots, Z_n)$. Let $\mathbb{E}_{\mathbf{Z}}$ denote conditional expectation given $\mathbf{Z}$. By Hoeffding's inequality, for all $L > 0$ and all i and j,

$$\mathbb{E}_{\mathbf{Z}} \exp[Z_{i,j} \varepsilon_i / L] \leqslant \exp[Z_{i,j}^2 / (2L^2)].$$

Define

$$\zeta = \max_{1 \leqslant j \leqslant M} \left| \sum_{i=1}^{n} Z_{i,j} \varepsilon_i \right|.$$

In view of Jensen's inequality, and by the independence assumption,

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}} \zeta^m &\leqslant L^m \mathbb{E}_{\mathbf{Z}} \log^m \left\{ \exp\left[\zeta / L\right] - 1 + \mathrm{e}^{m-1} \right\} \\
&\leqslant L^m \log^m \left\{ \mathbb{E}_{\mathbf{Z}} \exp\left[\zeta / L\right] - 1 + \mathrm{e}^{m-1} \right\} \\
&\leqslant L^m \log^m \left\{ 2M \exp\left[\sum_{i=1}^{n} \max_{1 \leqslant j \leq M} Z_{i,j}^2 / (2L^2) \right] + \mathrm{e}^{m-1} - 1 \right\}
\end{aligned}$$

There exists a constant $c(m) \geqslant 2$ such that for any $M \geqslant 1$ we have $e^{m-1} - 1 \leqslant c(m)M$. This yields

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}} \zeta^m &\leqslant L^m \log^m \left\{ c(m) M \left(\exp\left[\sum_{i=1}^{n} \max_{1 \leqslant j \leq M} Z_{i,j}^2 / (2L^2) \right] + 1 \right) \right\} \\
&\leqslant L^m \left\{ \log(2c(m)M) + \frac{\sum_{i=1}^{n} \max_{1 \leqslant j \leqslant M} Z_{i,j}^2}{2L^2} \right\}^m.
\end{aligned}$$

Choosing

$$L = \sqrt{\frac{\sum_{i=1}^{n} \max_{1 \leqslant j \leqslant M} Z_{i,j}^2}{2 \log(2c(m)M)}}$$

gives

$$\mathbb{E}_{\mathbf{Z}} \max_{1\leqslant j\leq M}\left|\sum_{i=1}^{n} Z_{i,j}\varepsilon_i\right|^m \leqslant \left[2\log(2c(m)M)\sum_{i=1}^{n}\max_{1\leqslant j\leqslant M} Z_{i,j}^2\right]^{m/2}.$$

Hence,

$$\mathbb{E}\max_{1\leqslant j\leqslant M}\left|\sum_{i=1}^{n} Z_{i,j}\varepsilon_i\right|^m \leqslant \left[2\log(2c(m)M)\right]^{m/2}\mathbb{E}\left[\sum_{i=1}^{n}\max_{1\leqslant j\leqslant M} Z_{i,j}^2\right]^{m/2}.$$

Finally, we de-symmetrize:

$$\left(\mathbb{E}\max_{1\leqslant j\leqslant M}\left|\sum_{i=1}^{n}(Z_{i,j}-\mathbb{E}Z_{i,j})\right|^m\right)^{1/m} \leqslant 2\left(\mathbb{E}\max_{1\leqslant j\leqslant M}\left|\sum_{i=1}^{n} Z_{i,j}\varepsilon_i\right|^m\right)^{1/m}$$

□

5.7 Auxiliary results

Here we collect two auxiliary results which are used in the above analysis. The first result is a useful bound on the tail of the chi-square distribution.

Lemma 5.4. *Let χ_T^2 be a chi-square random variable with T degrees of freedom. Then*

$$\Pr(\chi_T^2 > T + x) \leq \exp\left(-\frac{1}{8}\min\left(x, \frac{x^2}{T}\right)\right)$$

for all $x > 0$.

Proof. By the Wallace inequality [105] we have

$$\Pr(\chi_T^2 > T + x) \leq \Pr(\mathcal{N} > z(x)),$$

where $\mathcal{N}$ is the standard normal random variable and $z(x) = \sqrt{x - T\log(1 + x/T)}$. The result now follows from inequalities $\Pr(\mathcal{N} > z(x)) \leq \exp(-z^2(x)/2)$ and

$$u - \log(1 + u) \geq \frac{u^2}{2(1 + u)} \geq \frac{1}{4}\min\left(u, u^2\right),\ \forall u > 0.$$

□

The next result is a version of Nemirovski's inequality (see [36], Corollary 2.4 page 5).

Lemma 5.5. *Let $Y_1, \ldots, Y_n \in \mathbb{R}^M$ be independent random vectors with zero means and finite variance, and let $M \geqslant 3$. Then*

$$\mathbb{E}\left[|\sum_{i=1}^{n} Y_i|_\infty^2\right] \leqslant (2e\log M - e)\sum_{i=1}^{n}\mathbb{E}\left[|Y_i|_\infty^2\right],$$

where $|\cdot|_\infty$ is the ℓ_∞ norm.

Chapter 6

Sparsity oracle inequalities for the generalized Dantzig Selector

We propose a generalized version of the Dantzig selector. We show that it satisfies sparsity oracle inequalities in prediction and estimation. We consider then the particular case of high-dimensional linear regression model selection with the Lipschitz continuous loss function and the quadratic loss. In these cases we derive the sup-norm convergence rate and the sign concentration property of the generalized and usual Dantzig Selector under a mutual coherence assumption on the dictionary.

6.1 Introduction

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ be a measurable space. We observe a set of n i.i.d. random pairs $Z_i = (X_i, Y_i)$, $i = 1, \ldots, n$ where $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$. Denote by P the joint distribution of (X_i, Y_i) on $\mathcal{X} \times \mathcal{Y}$, and by P^X the marginal distribution of X_i. Let $Z = (X, Y)$ be a random pair in $\mathcal{Z}$ distributed according to P. For any real-valued function g on $\mathcal{X}$, define $||g||_\infty = \operatorname{ess\,sup}_{x \in \mathcal{X}} |g(x)|$, $\|g\| = \left(\int_{\mathcal{X}} g(x)^2 P^X(dx)\right)^{1/2}$ and $||g||_n = \left(\frac{1}{n}\sum_{i=1}^n g(X_i)^2\right)^{1/2}$. Let $\mathcal{D} = \{f_1, \ldots, f_M\}$ be a set of real-valued functions on $\mathcal{X}$ called the dictionary where $M \geqslant 2$. We assume that the functions of the dictionary are normalized, so that $\|f_j\| = 1$ for all $j = 1, \ldots, M$. We also assume that $||f_j||_\infty \leqslant L$ for some $L > 0$. For any $\theta \in \mathbb{R}^M$, define $f_\theta = \sum_{j=1}^M \theta_j f_j$ and $J(\theta) = \{j \ : \ \theta_j \neq 0\}$. Let $M(\theta) = |J(\theta)|$ be the cardinality of $J(\theta)$ and $\overrightarrow{\operatorname{sign}}(\theta) = (\operatorname{sign}(\theta_1), \ldots, \operatorname{sign}(\theta_M))^T$ where

$$\operatorname{sign}(t) = \begin{cases} 1 & \text{if } t > 0, \\ 0 & \text{if } t = 0, \\ -1 & \text{if } t < 0. \end{cases}$$

For any vector $\theta \in \mathbb{R}^M$ and any subset J of $\{1, \ldots, M\}$, we denote by θ_J the vector in $\mathbb{R}^M$ which has the same coordinates as θ on J and zero coordinates on the complement J^c of J. For any integers $1 \leqslant d, p < \infty$ and $w = (w_1, \ldots, w_d) \in \mathbb{R}^d$, the l_p norm of the vector w is denoted by $|w|_p \triangleq \left(\sum_{j=1}^d |w_j|^p\right)^{1/p}$, and $|w|_\infty \triangleq \max_{1 \leqslant j \leqslant d} |w_j|$.

Consider a function $\gamma : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}^+$ such that for any y in $\mathcal{Y}$ and u, u' in $\mathbb{R}$ we have

$$|\gamma(y, u) - \gamma(y, u')| \leqslant |u - u'|.$$

We assume furthermore that $\gamma(y, \cdot)$ is convex and differentiable for any $y \in \mathcal{Y}$. We assume that for any $y \in \mathcal{Y}$ the derivative $\partial_u \gamma(y, \cdot)$ is absolutely continuous. Then $\partial_u \gamma(y, \cdot)$ admits a derivative almost everywhere which we denote by $\partial_u^2 \gamma(y, \cdot)$. Consider the loss function $Q : \mathcal{Z} \times \mathbb{R}^M \to \mathbb{R}^+$ defined by

$$Q(z, \theta) = \gamma(y, f_\theta(x)). \tag{6.1}$$

The expected and empirical risk measures at point θ in $\mathbb{R}^M$ are defined respectively by

$$R(\theta) \triangleq \mathbb{E}\left(Q(Z, \theta)\right),$$

where $\mathbb{E}$ is the expectation sign, and

$$\hat{R}_n(\theta) \triangleq \frac{1}{n} \sum_{i=1}^n Q(Z_i, \theta).$$

Define the target vector as a minimizer of $R(\cdot)$ over $\mathbb{R}^M$:

$$\theta^* \stackrel{\triangle}{=} \arg\min_{\theta\in\mathbb{R}^M} R(\theta).$$

Note that the target vector is not necessarily unique. From now on, we assume that there exists a s-sparse solution θ^*, i.e., a solution with $M(\theta^*) \leqslant s$, and that this sparse solution is unique. We will see that this is indeed the case under the coherence condition on the dictionary (cf. Section 6.4 below).

Define the excess risk of the vector θ by

$$\mathcal{E}(\theta) = R(\theta) - R(\theta^*),$$

and its empirical version by

$$\mathcal{E}_n(\theta) = R_n(\theta) - R_n(\theta^*).$$

Our goal is to derive sparsity oracle inequalities for the excess risk and for the risk of θ^* in the l_1 norm and in the sup-norm.

We consider the following minimization problem:

$$\min_{\theta\in\Theta} |\theta|_1 \text{ subject to } \left|\nabla\hat{R}_n(\theta)\right|_\infty \leqslant r, \tag{6.2}$$

where $\nabla\hat{R}_n \stackrel{\triangle}{=} (\partial_{\theta_1}\hat{R}_n, \ldots, \partial_{\theta_M}\hat{R}_n)^T$, $r > 0$ is a tuning parameter defined later and Θ is a convex subset of $\mathbb{R}^M$ specified later. Solutions of (6.2), if they exist, will be taken as estimators of θ^*. Note that we will prove in Lemma 6.3 below that under Assumption 6.2 the set $\{\theta \in \Theta : \left|\nabla\hat{R}_n(\theta)\right|_\infty \leqslant r\}$ is non-empty with probability close to one. Then $\hat{\Theta}$ the set of all solutions of (6.2) is non-empty with probability close to one since the objective function in (6.2) is coercive.

The definition of our estimator (6.2) can be motivated as follows. Since the loss function $Q(z,\cdot)$ is convex and differentiable for any fixed $z \in \mathcal{Z}$, the expected risk R is also a convex function of θ and it is differentiable under mild conditions. Thus, minimizing R is equivalent to finding the zeros of ∇R. The quantity $\nabla\hat{R}_n(\theta)$ is the empirical version of $\nabla R(\theta)$. We choose the constant r such that the vector θ^* satisfies the constraint $|\nabla\hat{R}_n(\theta^*)| \leqslant r$ with probability close to 1. Then among all the vectors satisfying this constraint, we choose those with minimum l_1 norm. Note that if we consider the linear regression problem with the quadratic loss, we recognize in (6.2) the Dantzig minimization problem of Candes and Tao [16]. From now on, we will call (6.2) the generalized Dantzig minimization problem.

One may argue that considering general loss function, instead of the quadratic loss, we loose the computational tractability of the original Dantzig Selector. However, in the

applications considered in Section 6.4, the estimators can be computed efficiently. Indeed, concerning the logistic regression model, James and Radchenko [55] propose a modification of LARS to compute a solution of (6.2) and, even more, the whole solution path. For the regression model with a Lipschitz continuous loss function Φ such that $\Phi^{(2)} > 0$, a solution can be computed by a similar modification of the LARS algorithm.

Note that [55] considers only the logistic regression model and does not contain analysis of statistical properties of (6.2). In this chapter, we derive theoretical prediction, estimation and sign concentration results for the generalized Dantzig selector in a general setup including the logistic regression model. We also note that some of our results, for example the sign concentration property in random design setting are new even for the usual Dantzig Selector with random X_i and bounded noise are new (cf. Section 6.7).

Previous work on the Dantzig selector with the quadratic loss is due to Bickel et al. [6], Candes and Tao [16] and Koltchinskii [62, 63]. They proved sparsity oracle inequalities on the excess risk and for the estimation of θ^* for the l_p norm with $1 \leqslant p \leqslant 2$.

The problem (6.2) is closely related to the minimization problem:

$$\min_{\theta \in \Theta} \hat{R}_n(\theta) + r|\theta|_1, \tag{6.3}$$

which is a generalized version of the Lasso. For the Lasso estimator, Bunea et al [15] proved similar results in high-dimensional regression problems with the quadratic loss under a mutual coherence assumption [34] and Bickel et al [6] under a weaker Restricted Eigenvalue assumption. Koltchinskii [62] derived similar results for the Lasso in the context of high-dimensional regresssion with twice differentiable Lipschitz continuous loss functions under a restricted isometry assumption. Van de Geer [99, 100] obtained similar results for the Lasso in the context of generalized linear models with Lipschitz continuous loss functions. Wegkamp [107] analyzed the Lasso type estimators under hinge loss in classification. Lounici [71] derived sup-norm convergence rates and sign consistency of the Lasso and Dantzig estimators in a high-dimensional linear regression model with the quadratic loss. The techniques of our proofs are close to those in [6, 15, 62, 71, 99, 107].

The chapter is organized as follows. In Section 6.3 we derive sparsity oracle inequalities for the excess risk and for estimation of θ^* for the generalized Dantzig estimators defined by (6.2) in a stochastic optimization framework. In Section 6.4 we apply the results of Section 6.3 to the linear regression model with Lipschitz continuous loss and to the logistic regression model. In Section 6.5 we prove the sup-norm estimation consistency with rates under a mutual coherence assumption for the linear regression model with Lipschitz continuous loss.

In Section 6.6 we show a sign concentration property of the thresholded generalized Dantzig estimators for the linear regression model. In Section 6.7 we prove similar results as in the previous sections for the linear regression model with the quadratic loss and the usual Dantzig Selector.

6.2 Sparsity oracle inequalities for prediction and estimation with the l_1 norm

We need an assumption on the dictionary to derive prediction and estimation results for the generalized Dantzig estimators. We first state the Restricted Eigenvalue assumption [6].

Assumption 6.1.

$$\zeta(s) \stackrel{\triangle}{=} \min_{J_0 \subset \{1,\dots,M\}:|J_0| \leqslant s} \quad \min_{\Delta \neq 0:|\Delta_{J_0^c}|_1 \leqslant |\Delta_{J_0}|_1} \frac{||f_\Delta||}{|\Delta_{J_0}|_2} > 0.$$

It implies an "equivalence" between the two norms $|\Delta|_2$ and $\|f_\Delta\|$ on the subset $\{\Delta \neq 0 : |\Delta_{J(\Delta)^c}|_1 \leqslant |\Delta_{J(\Delta)}|_1\}$ of $\mathbb{R}^M$.

We need the following assumption on $\|f_{\theta^*}\|_\infty$.

Assumption 6.2. *There exists a constant $K > 0$ such that $\|f_{\theta^*}\|_\infty \leqslant K$.*

From now on we take for Θ the set

$$\Theta = \{\theta \in \mathbb{R}^M \, : \, \|f_\theta\|_\infty \leqslant K\}.$$

The following assumption is a version of the margin condition (cf. [96]). It links the excess risk to the functional norm $\|\cdot\|$.

Assumption 6.3. *For any $\theta \in \Theta$ there exits a constant $c > 0$ depending possibly on K such that*

$$\|f_\theta - f_{\theta^*}\| \leqslant c(R(\theta) - R(\theta^*))^{1/\kappa},$$

where $1 < \kappa \leqslant 2$.

We will prove in Section 6.4 below that this condition is always satisfied with the constant $\kappa = 2$ for the regression model with any Lipschitz continuous loss satisfying $\Phi^{(2)} > 0$ and for the logistic regression model. We also need the following technical assumption.

Assumption 6.4. *The constants K and L satisfy*

$$1 \leqslant K, L \leqslant \sqrt{\frac{n}{\log M}}.$$

Define the quantity

$$r = 24\sqrt{2}L\frac{\log M}{n} + 12\sqrt{6}\sqrt{\frac{\log M}{n}}. \tag{6.4}$$

We assume from now on that $r \leqslant 1$. Note that we take the parameter r of the above form such that θ^* satisfies the constraint $|\nabla R_n(\theta^*)| \leqslant r$ with overwhelming probability.

The main results of this section are the following sparsity oracle inequalities for the excess risk and for estimation of θ^* in the l_1 norm.

Theorem 6.1. *Let Assumptions 6.1 - 6.4 be satisfied. Take r as in (6.4). Assume that $M(\theta^*) \leqslant s$. Then, with probability at least $1 - M^{-1} - M^{-K} - 3M^{-2K}\log\frac{n}{\log M}$, we have*

$$\sup_{\hat{\theta}\in\hat{\Theta}} \mathcal{E}(\hat{\theta}) \leqslant \left(\frac{2(1+2K)cr\sqrt{s}}{\zeta(s)}\right)^{\frac{\kappa}{\kappa-1}} + \frac{\kappa}{3(\kappa-1)}r^2, \tag{6.5}$$

and

$$\begin{aligned}\sup_{\hat{\theta}\in\hat{\Theta}} |\hat{\theta} - \theta^*|_1 \leqslant & \left(\frac{2c\sqrt{s}}{\zeta(s)}\right)^{\frac{\kappa}{\kappa-1}} ((1+2K)r)^{\frac{1}{\kappa-1}} \\ & + \frac{K}{3(\kappa-1)(1+2K)}r.\end{aligned} \tag{6.6}$$

Note that the regularization parameter r does not depend on the variance of the noise if we consider the regression model with non-quadratic loss. In this case, the use of Lipschitz losses enables us to treat cases where the noise variable does not admit a finite second moment, e.g., the Cauchy distribution. The price to pay is that we need to assume that $\|f_{\theta^*}\|_\infty \leqslant K$ with known K.

Proof. For any $\hat{\theta} \in \hat{\Theta}$ define $\Delta = \hat{\theta} - \theta^*$. Set $\tilde{r} = r/6$. We have

$$\begin{aligned}\mathcal{E}(\hat{\theta}) - \mathcal{E}_n(\hat{\theta}) &= \frac{\mathcal{E}(\hat{\theta}) - \mathcal{E}_n(\hat{\theta})}{|\Delta|_1 + \tilde{r}}(|\Delta|_1 + \tilde{r}) \\ &\leqslant \sup_{\theta\in\Theta:\theta\neq\theta^*}\left(\frac{\mathcal{E}(\theta) - \mathcal{E}_n(\theta)}{|\theta - \theta^*|_1 + \tilde{r}}\right)(|\Delta|_1 + \tilde{r}).\end{aligned} \tag{6.7}$$

By Lemma 6.1 it holds on an event $\mathcal{A}_1$ of probability at least $1 - M^{-K} - 3M^{-2K}\log\frac{n}{\log M}$ that

$$\sup_{\theta\in\Theta:\theta\neq\theta^*} \frac{\mathcal{E}(\theta) - \mathcal{E}_n(\theta)}{|\theta - \theta^*|_1 + \tilde{r}} \leqslant 2Kr. \tag{6.8}$$

For any $\hat{\theta} \in \hat{\Theta}$, we have by definition of the Dantzig estimator that $|\hat{\theta}|_1 \leqslant |\theta^*|_1$. Thus

$$\begin{aligned} |\Delta_{J(\theta^*)^c}|_1 &= \sum_{j \in J(\theta^*)^c} |\hat{\theta}_j| \\ &\leqslant \sum_{j \in J(\theta^*)} |\theta^*_j| - |\hat{\theta}_j| \\ &\leqslant |\Delta_{J(\theta^*)}|_1. \end{aligned} \tag{6.9}$$

Define the function $g : t \to R_n(\theta^* + t\Delta)$. Clearly g is convex and differentiable on $[0,1]$. Thus, the function g' is nondecreasing on $[0,1]$ with derivative $g'(t) = \nabla R_n(\theta^* + t\Delta)^T \Delta$. The constraint $\left|\nabla \hat{R}_n(\theta)\right|_\infty \leqslant r$ in (6.2) and Lemma 6.3 yield, on an event $\mathcal{A}_2$ of probability at least $1 - M^{-1}$,

$$\begin{aligned} \mathcal{E}_n(\hat{\theta}) &= R_n(\hat{\theta}) - R_n(\theta^*) \\ &= \int_0^1 \nabla R_n(\theta^* + t\Delta)^T \Delta dt \\ &\leqslant r|\Delta|_1, \end{aligned} \tag{6.10}$$

for some numerical constant $C > 0$.

Combining (6.7)-(6.10) yields that on the event $\mathcal{A}_1 \cap \mathcal{A}_2$

$$\mathcal{E}(\hat{\theta}) \leqslant (2 + 4K) r |\Delta_{J(\theta^*)}|_1 + \frac{K}{3} r^2. \tag{6.11}$$

Set $C = 2(1 + 2K)$. We have

$$\begin{aligned} Cr|\Delta_{J(\theta^*)}|_1 &\leqslant Cr\sqrt{s}|\Delta_{J(\theta^*)}|_2 \\ &\leqslant \frac{Ccr\sqrt{s}}{\zeta(s)} \frac{\|f_\Delta\|}{c} \\ &\leqslant \frac{1}{\kappa'} \left(\frac{Ccr\sqrt{s}}{\zeta(s)} \right)^{\kappa'} + \frac{1}{\kappa} \left(\frac{\|f_\Delta\|}{c} \right)^{\kappa} \\ &\leqslant \frac{1}{\kappa'} \left(\frac{Ccr\sqrt{s}}{\zeta(s)} \right)^{\kappa'} + \frac{1}{\kappa} \mathcal{E}(\hat{\theta}^D), \end{aligned} \tag{6.12}$$

where we have used the Cauchy-Schwarz inequality in the first line, the inequality $xy \leqslant |x|^\kappa/\kappa + |y|^{\kappa'}/\kappa'$ that holds for any x, y in $\mathbb{R}$ and for any κ, κ' in $(1, \infty)$ such that $1/\kappa + 1/\kappa' = 1$ in the third line, and Assumption 2 in the last line. Combining (6.11) and (6.12) and the fact that $\tilde{r} \leqslant 1$ yields the first inequality. The second inequality is a consequence of (6.5) and (6.12). □

We state below intermediate results used in the proof of Theorem 6.1. These results can be proved using standard results of the theory of empirical processes. Lemma 6.1 proposes a bound on a the supremum of a weighted empirical process obtained via a peeling device technique. Lemma 6.2 is a standard Bernstein-type inequality on supremum of Rademacher averages. Lemma 6.3 uses similar techniques to bound the supremum of a gradient function.

Lemma 6.1. *Let Assumptions 6.2 and 6.4 be satisfied. Then, with probability at least* $1 - M^{-K} - 3M^{-2K} \log \frac{n}{\log M}$*, we have*

$$\sup_{\theta \in \Theta} \frac{|\mathcal{E}(\theta) - \mathcal{E}_n(\theta)|}{|\theta - \theta^*|_1 + r/6} \leqslant 2Kr, \tag{6.13}$$

where r *is defined in (6.4).*

Proof. For any $A > 0$, define the random variable

$$T_A = \sup_{\theta \in \Theta : |\theta - \theta^*|_1 \leqslant A} |\mathcal{E}_n(\theta) - \mathcal{E}(\theta)|.$$

For any θ in Θ and (x, y) in $\mathcal{Z}$ we have

$$|\gamma(y, f_\theta(x)) - \gamma(y, f_{\theta^*}(x))| \leqslant (L|\theta - \theta^*|_1) \wedge (2K),$$

and

$$\begin{aligned} \mathbb{E}\left(|\gamma(Y, f_\theta(X)) - \gamma(Y, f_{\theta^*}(X))|^2\right) &\leqslant \mathbb{E}(|f_\theta(X) - f_{\theta^*}(X)|^2) \\ &\leqslant ((\theta - \theta^*)^T G(\theta - \theta^*)) \wedge (4K^2) \\ &\leqslant (2|\theta - \theta^*|_1^2) \wedge (4K^2), \end{aligned}$$

since for any $\theta \in \mathbb{R}^M$

$$\begin{aligned} \theta^T G \theta &= |\theta|_2^2 + \sum_{j \neq k} \theta_j \theta_k G_{j,k} \theta \\ &\leqslant |\theta|_2^2 + \sum_{j \neq k} |\theta_j||\theta_k| \\ &\leqslant |\theta|_2^2 + |\theta|_1^2 \\ &\leqslant 2|\theta|_1^2, \end{aligned}$$

where we have used that $\|f_j\| = 1$, $\forall j$ and $|G_{j,k}| = |\mathbb{E}(f_j(X) f_k(X))| \leqslant 1$ in the second line.

We consider the quantity $\mathbb{E}(T_A)$. By standard symmetrization and contraction arguments (cf. [101] p. 108 and [66] p. 95) we obtain

$$\mathbb{E}(T_A) \leqslant 4\mathbb{E}\left(\sup_{\theta \in \Theta \,:\, |\theta - \theta^*|_1 \leqslant A} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f_{\theta - \theta^*}(X_i) \right| \right).$$

Then, observe that the mapping $u \to \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f_u(X_i)$ is linear, thus its supremum on a simplex is attained at one of its vertices. This yields

$$\mathbb{E}(T_A) \leqslant 4A\mathbb{E}\left(\max_{1\leqslant j\leqslant M}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_j(X_i)\right|\right).$$

Combining Assumption 6.4 and Lemma 6.2 we obtain

$$\mathbb{E}(T_A) \leqslant \frac{2}{3}Ar.$$

Assumption 6.2 and Bousquet's concentration inequality (cf. Theorem 2.3 in [11]) with $x = (A \vee 2K)\log M$, $c = 2(AL \wedge 2K)$ and $\sigma = \sqrt{2}(A \wedge 2K)$ yield

$$\mathbb{P}\left(T_A \geqslant \mathbb{E}(T_A) + \sqrt{\frac{2x}{n}}v + \frac{cx}{3n}\right) \leqslant M^{-(2K)\vee A},$$

where $v = \sigma^2 + 2c\mathbb{E}(T_A)$. By Assumption 6.4 and the definition of r we get

$$\begin{aligned}\sqrt{\frac{2x}{n}}v + \frac{cx}{3n} &\leqslant \frac{AK}{3}\left(2\sqrt{42}\sqrt{\frac{\log M}{n}} + 4L\frac{\log M}{n}\right)\\ &\leqslant \frac{AK}{3}r.\end{aligned}$$

Thus we get

$$\mathbb{P}\left(T_A \geqslant AKr\right) \leqslant M^{-(2K)\vee A}. \tag{6.14}$$

Recall that $\tilde{r} = r/6$. Define the following subsets of Θ

$$\begin{aligned}\Theta(I) &= \left\{\theta \in \Theta : |\theta - \theta^*|_1 \leqslant \tilde{r}\right\},\\ \Theta(II) &= \left\{\theta \in \Theta : \tilde{r} < |\theta - \theta^*|_1 \leqslant 2K\right\},\\ \Theta(III) &= \left\{\theta \in \Theta : |\theta - \theta^*|_1 > 2K\right\}.\end{aligned}$$

For any $t > 0$ define the probabilities

$$\begin{aligned}P_I &= \mathbb{P}\left(\sup_{\theta\in\Theta(I)}\frac{|\mathcal{E}(\theta) - \mathcal{E}_n(\theta)|}{|\theta - \theta^*|_1 + \tilde{r}} \geqslant t\right)\\ P_{II} &= \mathbb{P}\left(\sup_{\theta\in\Theta(II)}\frac{|\mathcal{E}(\theta) - \mathcal{E}_n(\theta)|}{|\theta - \theta^*|_1 + \tilde{r}} \geqslant t\right)\\ P_{III} &= \mathbb{P}\left(\sup_{\theta\in\Theta(III)}\frac{|\mathcal{E}(\theta) - \mathcal{E}_n(\theta)|}{|\theta - \theta^*|_1 + \tilde{r}} \geqslant t\right)\end{aligned}$$

For any $t > 0$ we have

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\frac{|\mathcal{E}(\theta)-\mathcal{E}_n(\theta)}{|\theta-\theta^*|_1+\tilde{r}}\geqslant t\right)\leqslant P_I+P_{II}+P_{III}.$$

Now, we bound from above the three probabilities on the right hand side of the above expression. Take $t = 2Kr$. Applying (6.14) to P_I yields that

$$P_I\leqslant\mathbb{P}\left(T_{\tilde{r}}\geqslant Kr^2\right)\leqslant M^{-2K},$$

since we have $\tilde{r}\leqslant\leqslant 1\leqslant K$ by Assumption 6.4.

Consider now P_{II}. We have

$$\Theta(II)\subset\bigcup_{j=0}^{j_0}\left\{\theta\in\Theta\,:\,A_{j+1}\leqslant|\theta-\theta^*|_1\leqslant A_j\right\},$$

where $A_j = 2^{1-j}K$, $j = 0,\ldots,j_0$ and j_0 is chosen such that $2^{1-j_0}K > \tilde{r}$ and $2^{-j_0}K\leqslant\tilde{r}$. Thus

$$\begin{aligned}
P_{II} &\leqslant \sum_{j=0}^{j_0}\mathbb{P}\left(T_{A_j}\geqslant 2A_{j+1}Kr\right)\\
&\leqslant \sum_{j=0}^{j_0}\mathbb{P}\left(T_{A_j}\geqslant A_jKr\right)\\
&\leqslant (j_0+1)M^{-2K}\\
&\leqslant \left(3\left(\log\frac{n}{\log M}\right)-1\right)M^{-2K}.
\end{aligned}$$

Consider finally P_{III}. We have

$$\Theta(III)\subset\bigcup_{j=0}^{\infty}\left\{\theta\in\Theta\,:\,\overline{A}_{j-1}<|\theta-\theta^*|_1\leqslant\overline{A}_j\right\},$$

where $\overline{A}_j = 2^{1+j}K$, $j\geqslant 0$. Thus

$$\begin{aligned}
P_{III} &\leqslant \sum_{j=1}^{\infty}\mathbb{P}\left(T_{\overline{A}_j}\geqslant 2\overline{A}_{j-1}Kr\right)\\
&\leqslant \sum_{j=0}^{j_0}\mathbb{P}\left(T_{A_j}\geqslant\overline{A}_jKr\right)\\
&\leqslant \sum_{j=1}^{\infty}M^{-\overline{A}_j}\\
&\leqslant M^{-K}.
\end{aligned}$$

□

We now study the quantity $\mathbb{E}\left(\max_{1\leqslant j\leqslant M}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_j(X_i)\right|\right)$. This is done in the next lemma.

Lemma 6.2. *We have*

$$\mathbb{E}\left(\max_{1\leqslant j\leqslant M}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f_j(X_i)\right|\right) \leqslant \tilde{r}, \tag{6.15}$$

where $\tilde{r}$ is defined in (6.4).

Proof. Define the random variables

$$U_j = \frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f_j(X_i).$$

The Bernstein inequality yields, for any $j = 1, \ldots, M$ and $t > 0$,

$$\mathbb{P}\left(|U_j| \geqslant t\right) \leqslant \exp\left(-\frac{t^2}{2(t\|f_j\|_\infty/(3\sqrt{n}) + \|f_j\|^2)}\right). \tag{6.16}$$

Set $b_j = \|f_j\|_\infty/(3\sqrt{n})$. Define the random variables $T_j = U_j \mathbb{1}_{|Y_j|>\|f_j\|^2/b_j}$ and $T_j' = U_j \mathbb{1}_{|Y_j|\leqslant\|f_j\|^2/b_j}$. For all $t > 0$ we have

$$\mathbb{P}\left(|T_j| > t\right) \leqslant 2\exp\left(-\frac{t}{4b_j}\right), \quad \mathbb{P}\left(|T_j'| > t\right) \leqslant 2\exp\left(-\frac{t^2}{4\|f_j\|^2}\right).$$

Define the function $h_\nu(x) = \exp(x^\nu) - 1$, where $\nu > 0$. This function is clearly convex for any $\nu > 0$. We have

$$\mathbb{E}\left(h_1\left(\frac{|T_j|}{12b_j}\right)\right) = \int_0^\infty e^t \mathbb{P}(|T_j| > 12b_j t)dt \leqslant 1,$$

where we have used Fubini's Theorem in the first equality. Since the function h_1 is convex and nonnegative, we have

$$\begin{aligned} h_1\left(\mathbb{E}\left(\max_{1\leqslant j\leqslant M}\frac{|T_j|}{12b_j}\right)\right) &\leqslant \mathbb{E}\left(h_1\left(\max_{1\leqslant j\leqslant M}\frac{|T_j|}{12b_j}\right)\right) \\ &\leqslant \mathbb{E}\left(\sum_{j=1}^M h_1\left(\frac{|T_j|}{12b_j}\right)\right) \\ &\leqslant M, \end{aligned}$$

where we have used the Jensen inequality. Since the function $h_1^{-1}(x) = \log(1+x)$ is increasing, we have

$$\begin{aligned} \mathbb{E}\left(\max_{1\leqslant j\leqslant M}\frac{|T_j|}{12b_j}\right) &\leqslant \log(1+M) \\ \mathbb{E}\left(\max_{1\leqslant j\leqslant M}|T_j|\right) &\leqslant 4\frac{\log(1+M)}{\sqrt{n}}\max_{1\leqslant j\leqslant M}\|f_j\|_\infty. \end{aligned} \tag{6.17}$$

Applying the same argument to the function h_2, we prove that

$$\mathbb{E}\left(\max_{1\leqslant j\leqslant M}|T_j'|\right)\leqslant 2\sqrt{3}\sqrt{\log(1+M)}\max_{1\leqslant j\leqslant M}\|f_j\|. \tag{6.18}$$

Combining (6.17) and (6.18) yields the result.

□

Lemma 6.3. *Let Assumptions 6.2 and 6.4 be satisfied. Then, with probability at least* $1-M^{-1}$*, we have*

$$|\nabla\hat{R}_n(\theta^*)|_\infty\leqslant r,$$

where r *is defined in Theorem 6.1.*

Proof. For any $1\leqslant j\leqslant M$ define

$$Z_j=\frac{1}{n}\sum_{i=1}^n\partial_u\gamma(Y_i,f_{\theta^*}(X_i))f_j(X_i).$$

Since the function $\theta\to\gamma(y,f_\theta(x))$ is differentiable w.r.t. θ and $|\partial_u\gamma(y,f_\theta(x))f_j(x)|\leqslant\|\partial_u\gamma\|_\infty L$ for any $(x,y)\in\mathcal{X}\times\mathcal{Y}$ and $\theta\in\mathbb{R}^M$, we have

$$\mathbb{E}(Z_j)=\frac{\partial R(\theta^*)}{\partial\theta_j}=0.$$

Next, similarly as in Lemmas 6.1 and 6.2, we prove that

$$\mathbb{E}(|\nabla\hat{R}_n(\theta^*)|_\infty)\leqslant 4\|\partial_u\gamma\|_\infty\tilde{r}.$$

Finally Bousquet's concentration inequality (cf. Theorem 6.7 in Section 6.7 below) yields that, with probability at least $1-M^{-1}$,

$$\begin{aligned}|\nabla\hat{R}_n(\theta^*)|_\infty&\leqslant\mathbb{E}(|\nabla\hat{R}_n(\theta^*)|_\infty)\\&\quad+\sqrt{2\frac{\log M}{n}\left(\|\partial_u\gamma\|_\infty^2+2\|\partial_u\gamma\|_\infty L\mathbb{E}(|\nabla\hat{R}_n(\theta^*)|_\infty)\right)}\\&\quad+\frac{\|\partial_u\gamma\|_\infty L\log M}{3n}\\&\leqslant 6\|\partial_u\gamma\|_\infty\tilde{r}.\end{aligned}$$

□

6.3 Examples

6.3.1 Robust regression with Lipschitz continuous loss

We consider the linear regression model

$$Y = f_{\theta^*}(X) + W, \tag{6.19}$$

where $X \in \mathbb{R}^d$ is a random vector, $W \in \mathbb{R}$ is a random variable independent of X whose distribution is symmetric w.r.t. 0 and $\theta^* \in \mathbb{R}^M$ is the unknown vector of parameters. Assume that the function ϕ is Lipschitz continuous and $\phi^{(2)} > 0$. Define $\tau(R) = \inf_{|u| \leqslant R} \phi^{(2)}(u)$. The loss function is defined by

$$Q(z, \theta) = \phi(y - f_\theta(x)), \tag{6.20}$$

where $z = (x, y) \in \mathbb{R}^d \times \mathbb{R}$ and $\theta \in \Theta$.

In the following lemma we prove that for this loss function Assumption 6.3 is satisfied with $\kappa = 2$ and $c = \mathbb{P}(|W| \leqslant \alpha)\tau(2K + \alpha)$ where the constant $\alpha > 0$ is chosen such that $c > 0$.

Lemma 6.4. *Let Q be defined by (6.20). Then for any $\theta \in \Theta$ we have*

$$\frac{\mathbb{P}(|W| \leqslant \alpha)\tau(2K + \alpha)}{2} \|f_\theta - f_{\theta^*}\|^2 \leqslant \mathcal{E}(\theta).$$

Proof. Set $\Delta = \theta - \theta^*$. Since ϕ' is absolutely continuous, we have for any $\theta \in \Theta$

$$\begin{aligned} Q(Z, \theta) - Q(Z, \theta^*) &= \phi'(W) f_{-\Delta}(X) \\ &\quad + \left[\int_0^1 \phi^{(2)}(W + t f_{-\Delta}(X))(1 - t) dt \right] f_\Delta(X)^2 \\ &\geqslant \phi'(W) f_{-\Delta}(X) + \frac{1}{2} \mathbb{P}(|W| \leqslant \alpha)\tau(2K + \alpha) f_\Delta(X)^2, \end{aligned}$$

since $\|f_\theta\|_\infty \leqslant K$ for any $\theta \in \Theta$. Taking the expectations we get

$$R(\theta) - R(\theta^*) \geqslant \frac{\mathbb{P}(|W| \leqslant \alpha)\tau(2K + \alpha)}{2} \|f_\Delta\|^2,$$

for any $\alpha > 0$ since ϕ' is odd and the distribution of W is symmetric w.r.t. 0. □

We have the following corollary of Theorem 6.1.

Corollary 6.1. *Let Assumptions 6.1, 6.2 and 6.4 be satisfied. If $M(\theta^*) \leqslant s$, then, with probability at least $1 - M^{-1} - M^{-K} - 3M^{-2K} \log \frac{n}{\log M}$, we have*

$$\sup_{\hat{\theta} \in \hat{\Theta}} \mathcal{E}(\hat{\theta}) \leqslant \frac{8(1 + 2K)^2}{\mathbb{P}(|W| \leqslant \alpha)\tau(2K + \alpha)\zeta(s)^2} s r^2 + \frac{2}{3} r^2,$$

and

$$\sup_{\hat{\theta}\in\hat{\Theta}} |\hat{\theta} - \theta^*|_1 \leqslant \frac{8(1+2K)}{\mathbb{P}(|W| \leqslant \alpha)\tau(2K+\alpha)\zeta(s)^2} sr + \frac{K}{3(1+2K)} r.$$

6.3.2 Logistic regression and similar models

We consider $Z = (X, Y) \in \mathcal{X} \times \{0, 1\}$ where $\mathcal{X}$ is a Borel subset of $\mathbb{R}^d$. The conditional probability $\mathbb{P}(Y = 1 \,|X = x) = \pi(x)$ is unknown where π is a function on $\mathcal{X}$ with values in $[0, 1]$. We assume that π is of the form

$$\pi(x) = \Phi'(f_{\theta^*}(x)), \tag{6.21}$$

where the function $\Phi : \mathbb{R} \to \mathbb{R}^*$ is convex, twice differentiable, of derivative Φ' with values in $[0, 1]$ and the vector $\theta^* \in \mathbb{R}^M$ is unknown. Consider, e.g., the logit loss function $\Phi(u) = \log(1 + e^u)$. We assume that Φ is known. Define the quantity

$$\tau(R) = \frac{1}{2} \inf_{|u|\leqslant R} \Phi^{(2)}(u), \tag{6.22}$$

for any $R \geqslant 0$. We want to estimate θ^* with the procedure (6.2) and the convex loss function

$$Q(z, \theta) = -y f_\theta(x) + \Phi(f_\theta(x)), \tag{6.23}$$

where $z = (x, y) \in \mathbb{R}^d \times \{0, 1\}$. Thus we need to check Assumption 6.3 to apply Theorem 6.1.

Lemma 6.5. *Let the loss function be of the form (6.23) where Φ satisfies the above assumptions. Then for any $\theta \in \mathbb{R}^M$ we have*

$$\tau(K)\|f_\theta - f_{\theta^*}\|^2 \leqslant \mathcal{E}(\theta).$$

Proof. For any $\theta \in \Theta$, we have

$$\begin{aligned} Q(Z, \theta) - Q(Z, \theta^*) &= \nabla Q(Z, \theta^*)^T(\theta - \theta^*) \\ &\quad + \left[\int_0^1 \Phi^{(2)}(H(X)^T(\theta^* + t(\theta - \theta^*)))(1-t)dt\right] f_\Delta(X)^2 \\ &\geqslant \nabla Q(Z, \theta^*)^T(\theta - \theta^*) + \tau(\|f_\theta\|_\infty \vee \|f_{\theta^*}\|_\infty) f_\Delta(X)^2. \end{aligned}$$

Since $\|\nabla Q(\cdot, \cdot)\|_\infty < \infty$, we can differentiate under the expectation sign, so that

$$\mathbb{E}(\nabla Q(Z, \theta^*)^T(\theta - \theta^*)) = \nabla R(\theta^*) = 0.$$

Thus

$$\mathcal{E}(\theta) \geqslant \tau(\|f_\theta\|_\infty \vee \|f_{\theta^*}\|_\infty)\|f_\theta - f_{\theta^*}\|^2.$$

□

Thus Assumption 6.3 is satisfied with the constants $\kappa = 2$ and $c = \frac{1}{\sqrt{\tau(K)}}$. We have the following corollary of Theorem 6.1.

Corollary 6.2. *Let Assumptions 6.1, 6.2 and 6.4 be satisfied. If $M(\theta^*) \leqslant s$, then, with probability at least $1 - M^{-1} - M^{-K} - 3M^{-2K} \log \frac{n}{\log M}$, we have*

$$\sup_{\hat{\theta} \in \hat{\Theta}} \mathcal{E}(\hat{\theta}) \leqslant \frac{4(1+2K)^2}{\tau(K)\zeta(s)^2} sr^2 + \frac{2}{3} r^2,$$

and

$$\sup_{\hat{\theta} \in \hat{\Theta}} |\hat{\theta} - \theta^*|_1 \leqslant \frac{4(1+2K)}{\tau(K)\zeta(s)^2} sr + \frac{K}{3(1+2K)} r.$$

6.4 Sup-norm convergence rate for the regression model with Lipschitz continuous loss

In this section, we derive the sup-norm convergence rate of the Dantzig estimators to the target vector θ^* in the linear regression model under a mutual coherence assumption on the dictionary and general Lipschitz continuous loss function ϕ such that $\phi^{(2)} > 0$ on $\mathbb{R}$. The proof relies on the fact that the Hessian matrix of the risk also satisfies the mutual coherence condition for this particular model. Unfortunately, we cannot proceed similarly in the general case because the Hessian matrix of the risk at point θ^* does not necessarily satisfy the mutual coherence condition even if the Gram matrix of the dictionary does.

Denote by $\Psi(\theta)$ the Hessian matrix of the risk R evaluated at θ. With our assumptions on the dictionary $\mathcal{D}$ and on the function ϕ, for any $\theta \in \mathbb{R}^M$ we have

$$\Psi(\theta) \triangleq \nabla^2 R(\theta) = \left(\mathbb{E}\left(\phi^{(2)}(Y, f_\theta(X)) f_j(X) f_k(X) \right) \right)_{1 \leqslant j,k \leqslant M}.$$

Note that for the quadratic loss we have $\Psi(\cdot) \equiv 2G$ where G is the Gram matrix of the design. For Lipschitz loss functions the Hessian matrix Ψ varies with θ.

We consider the linear regression model (6.19). For any functions $g, h : \mathcal{X} \to \mathbb{R}$, denote by $< g, h >$ the scalar product $\mathbb{E}(g(X)h(X))$. Define the Gram matrix G by

$$G = (< f_j, f_k >)_{1 \leqslant j,k \leqslant M}.$$

From now on, we assume that G satisfies a mutual coherence condition.

Assumption 6.5. *The Gram matrix $G = (< f_j, f_k >)_{1 \leqslant j,k \leqslant M}$ satisfies*

$$G_{j,j} = 1, \, \forall 1 \leqslant j \leqslant M,$$

and

$$\max_{j\neq k}|G_{j,k}| \leqslant \frac{1}{3\beta s},$$

where $s \geqslant 1$ is an integer and $\beta > 1$ is a constant.

This assumption is stronger than Assumption 6.1. We have indeed the following Lemma (cf. Lemma 2 in [71]).

Lemma 6.6. *Let Assumption 6.5 be satisfied. Then*

$$\zeta(s) \stackrel{\triangle}{=} \min_{J\subset\{1,\cdots,M\},|J|\leqslant s}\ \min_{\Delta\neq 0:|\Delta_{J^c}|_1\leqslant|\Delta_J|_1} \frac{\|f_\Delta\|}{|\Delta_J|_2} \geqslant \sqrt{1-\frac{1}{\beta}} > 0.$$

Note that Assumption 6.5 the vector θ^* satisfying (6.19) such that $M(\theta^*) \leqslant s$ is **unique**. Consider indeed two vectors θ^1 and θ^2 satisfying (6.19) such that $M(\theta^1) \leqslant s$ and $M(\theta^2) \leqslant s$. Denote $\theta = \theta^1 - \theta^2$ and $J = J(\theta^1) \cup J(\theta^2)$. Clearly we have $f_\theta(X) = 0$ a.s. and $M(\theta) \leqslant 2s$. Assume that θ^1 and θ^2 are distinct. Then,

$$\begin{aligned}
\frac{\|f_\theta\|_2^2}{|\theta|_2^2} &= 1 + \frac{\theta^T(G-I_M)\theta}{|\theta|_2^2} \\
&\geqslant 1 - \frac{1}{3\beta s}\sum_{i,j=1}^{M}\frac{|\theta_i||\theta_j|}{|\theta|_2^2} \\
&\geqslant 1 - \frac{1}{3\beta} > 0,
\end{aligned}$$

where we have used the Cauchy-Schwarz inequality. This contradicts the fact that $f_\theta(X) = 0$ a.s.

For the linear regression model, the Hessian matrix Ψ at point θ is

$$\Psi(\theta) = \mathbb{E}(\phi^{(2)}(f_{\theta^*-\theta}(X)+W)f_j(X)f_k(X))_{1\leqslant j,k\leqslant M}.$$

We observe that

$$\Psi(\theta^*) = \mathbb{E}\phi^{(2)}(W)G.$$

Thus $\Psi(\theta^*)$ satisfies a condition similar to Assumption 6.5 but with a different constant if $\mathbb{E}\phi^{(2)}(W) > 0$. The empirical Hessian matrix $\hat{\Psi}$ at point $\theta \in \mathbb{R}^M$ is defined by

$$\hat{\Psi}_{j,k}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\phi(f_{\theta^*-\theta}(X_i)+W_i)f_j(X_i)f_k(X_i),\ 1 \leqslant j,k \leqslant M.$$

We will prove that the empirical Hessian matrix $\hat{\Psi}(\theta)$ satisfies a mutual coherence condition for any θ in a small neighborhood of θ^* under some additional assumptions given below.

First, we need an additional mild assumption on ϕ.

Assumption 6.6. *The function ϕ is such that $\phi^{(2)}$ is Lipschitz continuous and bounded from above by* 1.

We impose a restriction on the sparsity s.

Assumption 6.7. *The sparsity s satisfies $s \leqslant \frac{1}{\sqrt{r}}$.*

This implies that we can recover the sparse vectors with at most $O\left((n/\log M)^{1/4}\right)$ nonzero components.

Define $V_\eta = \{\theta \in \Theta : |\theta - \theta^*|_1 \leqslant \eta\}$ where $\eta = C_1 rs$ and

$$C_1 = \frac{8(1+2K)\beta}{\mathbb{P}(|W| \leqslant \alpha)\tau(2K+\alpha)(\beta-1)} + \frac{1}{6}. \tag{6.24}$$

Consider the event

$$E = \left\{ \sup_{1 \leqslant j,k \leqslant M, \theta \in V_\eta} \left| \hat{\Psi}_{j,k}(\theta) - \Psi_{j,k}(\theta) \right| \leqslant 8L^3\eta + 4L\tilde{r} + \frac{C_2}{\sqrt{n}s^2} \right\}, \tag{6.25}$$

where

$$C_2 = 2\sqrt{1 + (1+L^2)\left(8C_1L^3 + \frac{4L}{s}\right)} + \frac{1+L^2}{3}.$$

We have the following intermediate result.

Lemma 6.7. *Let Assumptions 6.2- 6.6 be satisfied. Then* $\mathbb{P}(E) \geqslant 1 - \exp(-\sqrt{\log M})$.

Proof. Define the variable

$$Z = \sup_{1 \leqslant j,k \leqslant M, \theta \in V_\eta} \left| \hat{\Psi}_{j,k}(\theta) - \Psi_{j,k}(\theta) \right|.$$

Applying the Bousquet concentration inequality (cf. Theorem 6.7 in Section 6.7) with the constants $c = (1+L^2)/n$, $\sigma^2 = 2/n^2$ and $x = \frac{\sqrt{n}}{s^2}$ yields that, with probability at least $1 - e^{-x}$,

$$Z \;\leqslant\; \mathbb{E}(Z) + \frac{2}{\sqrt{n}s}\sqrt{1 + (1+L^2)\mathbb{E}(Z)} + \frac{1+L^2}{3\sqrt{n}s^2}. \tag{6.26}$$

We study now the quantity $\mathbb{E}(Z)$. A standard symmetrization and contraction argument yields

$$\begin{aligned}
\mathbb{E}(Z) &\leqslant 2\mathbb{E}\left(\sup_{1 \leqslant j,k \leqslant M, \theta \in V_\eta} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i \phi(f_{\theta^*-\theta}(X_i) + W_i) f_j(X_i) f_k(X_i) \right| \right) \\
&\leqslant 2\mathbb{E}\left(\left| \frac{1}{n}\sum_{i=1}^n \epsilon_i \phi^{(2)}(W_i) f_j(X_i) f_k(X_i) \right| \right) \\
&\quad + 2\mathbb{E}\left(\sup_{1 \leqslant j,k \leqslant M, \theta \in V_\eta} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i (\phi^{(2)}(f_{\theta^*-\theta}(X_i) + W_i) - \phi^{(2)}(W_i)) f_j(X_i) f_k(X_i) \right| \right). \quad (6.27)
\end{aligned}$$

Denote by (I) and (II) respectively the first term and the second term on the right hand side of the above expression. The contraction principle yields

$$(I) \leqslant 4\mathbb{E}\left(\max_{1\leqslant j,k\leqslant M}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_j(X_i)f_k(X_i)\right|\right). \tag{6.28}$$

Then, similarly as in the proof of Lemma 6.2 we get

$$\mathbb{E}\left(\max_{1\leqslant j,k\leqslant M}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_j(X_i)f_k(X_i)\right|\right) \leqslant L\tilde{r}.$$

For (II) we have

$$\begin{aligned}(II) &\leqslant 2L^2\mathbb{E}\left(\sup_{\theta\in V_\eta}\frac{1}{n}\sum_{i=1}^{n}|\phi^{(2)}(f_\Delta(X_i)+W_i)-\phi^{(2)}(W_i)|\right)\\ &\leqslant 8L^3\eta. \end{aligned}\tag{6.29}$$

Assumptions 6.4 and 6.7 yield that $s \leqslant \left(\frac{n}{\log M}\right)^{1/4}$. Combining (6.26)-(6.29) yields the result. □

We need an additional technical assumption.

Assumption 6.8. *We have* $9L^3\eta + 4L\tilde{r} + \frac{C_2}{\sqrt{n}s^2} \leqslant \frac{\mathbb{E}\phi^{(2)}(W)}{2}$.

This is a mild assumption. It is indeed satisfied for n large enough if we assume that $\mathbb{E}\phi^{(2)}(W) > 0$ since Assumption 6.4 implies that $r \to 0$ as $n \to \infty$.

We have the following result on the empirical Hessian matrix.

Lemma 6.8. *Let Assumptions 6.2-6.8 be satisfied. Then, with probability at least* $1 - \exp(-\sqrt{\log M})$*, for any* $\theta \in V_\eta$*, we have*

$$\begin{aligned}\min_{1\leqslant j\leqslant M}|\hat{\Psi}_{j,j}(\theta)| &\geqslant \frac{\mathbb{E}\phi^{(2)}(W)}{2},\\ \max_{j\neq k}|\hat{\Psi}_{j,k}(\theta)| &\leqslant \frac{C_3}{s},\end{aligned}\tag{6.30}$$

where $C_3 = \frac{1}{3\beta} + 9L^3C_1 + 4L\tilde{r} + \frac{C_2}{\sqrt{n}s}$.

Proof. For any θ in V_η and any j, k in $\{1, \ldots, M\}$ we have

$$\begin{aligned}|\Psi_{j,k}(\theta) - \Psi_{j,k}(\theta^*)| &= |\mathbb{E}\left((\phi^{(2)}(f_\Delta(X)+W) - \phi^{(2)}(W))f_j(X)f_k(X)\right)|\\ &\leqslant L^2\mathbb{E}|f_\Delta(X)|\end{aligned}$$

where $\Delta = \theta - \theta^*$.

For any $\theta \in V_\eta$ we have $|f_\Delta(X)| \leqslant L\eta$. Then

$$|\Psi_{j,k}(\theta) - \Psi_{j,k}(\theta^*)| \leqslant L^3\eta, \tag{6.31}$$

We have for any j, k, θ

$$\hat{\Psi}_{j,k}(\theta) = \Psi_{j,k}(\theta^*) + \hat{\Psi}_{j,k}(\theta) - \Psi_{j,k}(\theta) + \Psi_{j,k}(\theta) - \Psi_{j,k}(\theta^*).$$

Lemma 6.7 and (6.31) yield that, on the event E, for any $\theta \in V_\eta$,

$$\min_{1 \leqslant j \leqslant M} \hat{\Psi}_{j,j}(\theta) \geqslant \mathbb{E}\phi^{(2)}(W) - 9L^3\eta - 4L\tilde{r} - \frac{C_2}{\sqrt{n}s^2},$$

and

$$\max_{j \neq k} |\hat{\Psi}_{j,k}(\theta)| \leqslant \frac{C_3}{s}.$$

□

Now we can derive the optimal sup-norm convergence rate of the Dantzig estimators.

Theorem 6.2. *Let Assumptions 6.2-6.8 be satisfied. If $M(\theta^*) \leqslant s$, then, on an event of probability at least $1 - M^{-1} - M^{-K} - \exp(-\sqrt{\log M}) - 3M^{-2K}\log\frac{n}{\log M}$, we have*

$$\sup_{\hat{\theta} \in \hat{\Theta}} |\hat{\theta} - \theta^*|_\infty \leqslant C_4 r,$$

where r is defined in Theorem 6.1,

$$C_4 = \frac{4 + 2C_1C_3}{\mathbb{E}\phi^{(2)}(W)},$$

with C_1 and C_3 defined respectively in (6.24) and Lemma 6.8.

Proof. For any $\hat{\theta}$ in $\hat{\Theta}$ we have

$$\nabla R_n(\hat{\theta}) - \nabla R_n(\theta^*) = \left[\int_0^1 \hat{\Psi}(\theta^* + t\Delta)dt\right]\Delta,$$

where $\Delta = \hat{\theta} - \theta^*$.

The definition of our estimator, Lemma 6.3 and Corollary 6.1 yield that, on an event $\mathcal{A}_1$ of probability at least $1 - M^{-1} - \exp(-\sqrt{\log M}) - 3M^{-2K}\log\frac{n}{\log M}$, we have that $\hat{\theta} \in V_\eta$ and

$$\left|\left[\int_0^1 \hat{\Psi}(\theta^* + t\Delta)dt\right]\Delta\right|_\infty \leqslant 2r.$$

Lemma 6.8 yields that, on the event $\mathcal{A}_1 \cap E$,

$$\frac{\mathbb{P}(|W| \leqslant 2K + \alpha)}{2} |\Delta|_\infty \leqslant 2r + \frac{C_3}{s}|\Delta|_1,$$

so that

$$|\Delta|_\infty \leqslant C_4 r.$$

□

Note that Theorem 6.2 holds true for the Lasso estimators (2) with exactly the same proof, provided that a result similar to Theorem 6.1 is valid for the Lasso estimators. This is in fact the case (cf. [99, 62]).

6.5 Sign concentration property with Lipschitz continuous loss

Now we study the sign concentration property of the Dantzig estimators. We need an additional assumption on the magnitude of the nonzero components of θ^*.

Assumption 6.9. *We have*

$$\rho \stackrel{\Delta}{=} \min_{j \in J(\theta^*)} |\theta_j^*| > 2C_4 r,$$

where r is defined in Theorem 6.1 and C_4 is defined in Theorem 6.2.

We can find similar assumptions on ρ in the work on sign consistency of the Lasso estimator mentioned above. More precisely, the lower bound on ρ is of the order $(s(\log M)/n)^{1/4}$ in [82], $n^{-\delta/2}$ with $0 < \delta < 1$ in [104, 116], $\sqrt{(\log Mn)/n}$ in [12], $\sqrt{s(\log M)/n}$ in [114] and r in [71].

We introduce the following thresholded version of our estimator. For any $\hat{\theta} \in \hat{\Theta}$ the associated thresholded estimator $\tilde{\theta} \in \mathbb{R}^M$ is defined by

$$\tilde{\theta}_j = \begin{cases} \hat{\theta}_j, & \text{if } |\hat{\theta}_j| > C_4 r, \\ 0 & \text{elsewhere.} \end{cases} \tag{6.32}$$

Denote by $\tilde{\Theta}$ the set of all such $\tilde{\theta}$. We have first the following non-asymptotic result that we call sign concentration property.

Theorem 6.3. *Let Assumptions 6.2 and 6.5-6.9 be satisfied. If $M(\theta^*) \leqslant s$, then*

$$\mathbb{P}\left(\overrightarrow{\operatorname{sign}}(\tilde{\theta}) = \overrightarrow{\operatorname{sign}}(\theta^*),\ \forall \tilde{\theta} \in \tilde{\Theta}\right) \geqslant 1 - M^{-1} - M^{-K} - \exp(-\sqrt{\log M}) - 3M^{-2K} \log \frac{n}{\log M}.$$

Theorem 6.3 guarantees that the sign vector of every vector $\tilde{\theta} \in \tilde{\Theta}$ coincides with that of θ^* with probability close to one.

Proof. Theorem 6.2 yields $\sup_{\hat{\theta}\in\hat{\Theta}} |\hat{\theta} - \theta^*|_\infty \leqslant C_3 r$ on an event $\mathcal{A}$ of probability at least $1 - 6M^{-1}$. Take $\hat{\theta} \in \hat{\Theta}$. For $j \in J(\theta^*)^c$, we have $\theta_j^* = 0$, and $|\hat{\theta}_j| \leqslant c_2 r$ on $\mathcal{A}$. For $j \in J(\theta^*)$, we have $|\theta_j^*| \geqslant 2C_3 r$ and $|\theta_j^*| - |\hat{\theta}_j| \leqslant |\theta_j^* - \hat{\theta} C_j| \leqslant C_3 r$ on $\mathcal{A}$. Since we assume that $\rho > 2C_3$, we have on $\mathcal{A}$ that $|\hat{\theta}_j| \geqslant> c_2 r$. Thus on the event $\mathcal{A}$ we have: $j \in J(\theta^*) \Leftrightarrow |\hat{\theta}_j| > c_2 r$. This yields $\text{sign}(\tilde{\theta}_j) = \text{sign}(\hat{\theta}_j) = \text{sign}(\theta_j^*)$ if $j \in J(\theta^*)$ on the event $\mathcal{A}$. If $j \notin J(\theta^*)$, $\text{sign}(\theta_j^*) = 0$ and $\tilde{\theta}_j = 0$ on $\mathcal{A}$, so that $\text{sign}(\tilde{\theta}_j) = 0$. The same reasoning holds true simultaneously for all $\hat{\theta} \in \hat{\Theta}$ on the event $\mathcal{A}$. Thus, we get the result. □

6.6 Sup-norm estimation and sign concentration property with the quadratic loss

Consider the regression model of Section 6.4

$$Y_i = f_{\theta^*}(X_i) + W_i, \;\; 1 \leqslant i \leqslant M,$$

where the noise variables W_i are now assumed i.i.d. such that $\mathbb{E}(W_i) = 0$ and $|W_i| \leqslant \overline{K}$ a.s. for some $\overline{K} \geqslant 1$. Note that this condition is not necessary, we can prove similar results under a sub-gaussian assumption on W_i. Koltchinskii [63] considered the usual Dantzig Selector

$$\min_{\theta\in\mathbb{R}^M} \left\{ |\theta|_1 \; : \; \max_{1\leqslant j\leqslant M} \left| n^{-1} \sum_{i=1}^n (f_\theta(X_i) - Y_i) f_j(X_i) \right| \leqslant r \right\}, \tag{6.33}$$

where r is now defined as follows

$$r = \overline{K} \left(4\sqrt{3} \sqrt{\frac{\log M}{n}} + 8L \frac{\log M}{n} \right). \tag{6.34}$$

From now on we assume that $r \leqslant 1$. Denote by $\hat{\Theta}$ the set of solutions of (6.33).

Koltchinskii [63] derived in Theorem 7.5 p.134 a result similar to Theorem 6.1 in Section 6.3 above under the Assumption 6.1 on the dictionary and the following additional condition

$$\|f_\theta\|_{L_1(P^X)} \leqslant \|f_\theta\| \leqslant B \|f_\theta\|_{L_1(P^X)}, \; \forall \theta \in \mathbb{R}^M \tag{6.35}$$

for some constant $B > 0$, where for any function $g : \mathcal{X} \to \mathbb{R}$, $\|g\|_{L^1(P^X)} = \int_{\mathcal{X}} |g(x)| P^X(dx)$.

Theorem 6.4 (Koltchinskii [63] p. 134)**.** *Suppose condition (6.35) holds. Let Assumption 6.1 be satisfied. If $M(\theta^*) \leqslant s$, then with probability at least $1 - M^{-1}$, we have*

$$\sup_{\hat{\theta} \in \hat{\Theta}} \|f_{\hat{\theta}} - f_{\theta^*}\|^2 \leqslant C \frac{B^4}{\zeta(s)^2} s r^2 \tag{6.36}$$

$$\sup_{\hat{\theta} \in \hat{\Theta}} |\hat{\theta} - \theta^*|_2 \leqslant C \frac{B^2}{\zeta(s)^2} \sqrt{s} r, \tag{6.37}$$

where $C > 0$ is a numerical constant .

Condition (6.35) is always satisfied for any two norms on finite dimensional spaces for some constant $B > 0$. However, as pointed out in [63], the constant B may depend on M for arbitrary dictionary of functions $\mathcal{D} = \{f_1, \ldots, f_M\}$. As a consequence, the rates in (6.36) and (6.37) may be sub-optimal for arbitrary dictionary $\mathcal{D}$. Note that we do not use condition (6.35) to prove our results in Theorem 6.1. Consequently the rates derived in Section 6.4 for the regression model with Lipschitz continuous loss and the logistic regression model are optimal for any dictionary $\mathcal{D}$ satisfying Assumption 6.1. In fact the condition (6.35) is not necessary. We can indeed prove a result similar to Theorem 6.1 with the optimal rates for the regression model with random design and the quadratic loss for any dictionary $\mathcal{D}$ satisfying Assumption 6.1.

Recall that $G = (< f_j, f_k >)_{1 \leqslant j,k \leqslant M}$. Denote by

$$\hat{G} = \left(\frac{1}{n} \sum_{i=1}^{n} f_j(X_i) f_k(X_i) \right)_{1 \leqslant j,k \leqslant M}$$

the empirical Gram matrix.

Lemma 6.9. *Assume that $1 \leqslant L \leqslant r^{-1}$, where r is defined in (6.34). Then, with probability at least $1 - M^{-1}$,*

$$\max_{1 \leqslant j,k \leqslant M} |\hat{G}_{j,k} - G_{j,k}| \leqslant \frac{3L}{2} r.$$

Proof. We have

$$\|f_j(\cdot) f_k(\cdot) - < f_j, f_k > \| \leqslant 1 + (\mathbb{E}(f_j(X)^2 f_k(X)^2))^{1/2} \leqslant 1 + L,$$

and

$$\|f_j(\cdot) f_k(\cdot) - < f_j, f_k > \|_\infty \leqslant 1 + L^2,$$

since $\|f_j\| = 1$ and $\|f_j\|_\infty \leqslant L$, $\forall 1 \leqslant j \leqslant M$. Consider the random variable

$$Z = \max_{1 \leqslant j,k \leqslant M} |\hat{G}_{j,k} - G_{j,k}|.$$

A standard symmetrization argument yields

$$\mathbb{E}(Z) \leqslant 2\mathbb{E}\left(\max_{1\leqslant j,k\leqslant M} |n^{-1}\sum_{i=1}^{n} \epsilon_i f_j(X_i) f_k(X_i)|\right),$$

where $\epsilon_1, \ldots, \epsilon_n$ is a Rademacher sequence. Then similarly to the proof of Lemma 6.2, we get

$$\mathbb{E}(Z) \leqslant L\left(4\sqrt{3}\sqrt{\frac{\log M}{n}} + 8L\frac{\log M}{n}\right) \leqslant Lr.$$

Applying Bousquet's version of Talagrand's inequality to Z with $x = \log M$, $c = (1+L^2)$ and $\sigma = 1 + L$ yields, with probability at least $1 - M^{-1}$,

$$Z \;\leqslant\; \mathbb{E}(Z) + \sqrt{\frac{2x}{n}(\sigma^2 + c\mathbb{E}(Z))} + \frac{c}{3}\frac{\log M}{n} \leqslant \frac{3}{2}Lr.$$

□

We state an additional technical assumption.

Assumption 6.10. *We have* $Lrs \leqslant \frac{\zeta(s)^2}{12}$ *where* $\zeta(s)$ *is defined in Assumption 6.1.*

This assumption imposes the additional restriction $s = O(r^{-1})$ due to randomness of the design as compared to [71] where the deterministic design is considered and no such restriction on s is needed.

Theorem 6.5. *Let* r *be defined in (6.34). Let Assumptions 6.1 and 6.10 be satisfied. If* $M(\theta^*) \leqslant s$*, then with probability at least* $1 - 2M^{-1}$*,*

$$\sup_{\hat{\theta}\in\hat{\Theta}} \|f_{\hat{\theta}} - f_{\theta^*}\|^2 \;\leqslant\; 144\frac{sr^2}{\zeta(s)^2} \tag{6.38}$$

$$\sup_{\hat{\theta}\in\hat{\Theta}} \|\hat{\theta} - \theta^*\|_1 \;\leqslant\; 24\frac{sr}{\zeta(s)^2}. \tag{6.39}$$

Proof. For any $\hat{\theta} \in \hat{\Theta}$, set $\Delta = \hat{\theta} - \theta^*$ and $J^* = J(\theta^*)$. By definition of the Dantzig Selector we have

$$|\hat{G}\Delta|_\infty \leqslant r + Z,$$

where $Z = \max_{1\leqslant j\leqslant M} |n^{-1}\sum_{i=1}^{n} W_i f_j(X_i)|$. Similarly to the proof of Lemma 6.2 we get $\mathbb{E}(Z) \leqslant r$. Then, using Theorem 6.7 we get, with probability at least $1 - M^{-1}$

$$Z \leqslant \mathbb{E}(Z) + \sqrt{\frac{\log M}{n}(\overline{K}^2 + L\overline{K}\mathbb{E}(Z))} + \frac{L\overline{K}}{3}\frac{\log M}{n} \leqslant 2r.$$

Thus with the same probability $|\hat{G}\Delta|_\infty \leqslant 3r$. By Lemma 6.9 we have with probability at least $1 - 2M^{-1}$

$$\begin{aligned} \|f_\Delta\|^2 &= \Delta^T G \Delta \\ &= \Delta^T (G - \hat{G})\Delta + \Delta^T \hat{G}\Delta \\ &\leqslant \frac{3}{2} Lr|\Delta|_1^2 + 3r|\Delta|_1. \end{aligned}$$

We have

$$|\Delta|_2 \leqslant |\Delta|_1 \leqslant 2|\Delta_{J^*}|_1 \leqslant 2\sqrt{s}|\Delta_{J^*}|_2 \leqslant \frac{2\sqrt{s}}{\zeta(s)}\|f_\Delta\|,$$

where we have used that $|\Delta_{J^{*c}}|_1 \leqslant |\Delta_{J^*}|_1$ for the Dantzig Selector, the Cauchy-Schwarz inequality and Assumption 6.1. The last two displays and Assumption 6.10 yield

$$\frac{1}{2}\|f_\Delta\|^2 \leqslant \left(1 - 6\frac{Lrs}{\zeta(s)^2}\right)\|f_\Delta\|^2 \leqslant r|\Delta|_1 \leqslant \frac{2r\sqrt{s}}{\zeta(s)}\|f_\Delta\|.$$

This yields the results. □

We now state the main result of this section.

Theorem 6.6. *Let $s \geqslant 2$ and r be defined in (6.34). Let Assumptions 6.5 and 6.10 be satisfied. Define $c = \left(\frac{72}{7} + \frac{64}{7(\beta-1)}\right)$. If $M(\theta^*) \leqslant s$, then with probability at least $1 - 2M^{-1}$*

$$\sup_{\hat{\theta}\in\hat{\Theta}} |\hat{\theta} - \theta^*|_\infty \leqslant cr \tag{6.40}$$

In addition, if $\min_{j\in J(\theta^*)} |\theta_j^*| \geqslant 2cr$, then with the same probability for any $\tilde{\theta} \in \tilde{\Theta}$ defined as in (6.32) but with the threshold cr instead of $C_4 r$ we have

$$\overrightarrow{\text{sign}}(\tilde{\theta}) = \overrightarrow{\text{sign}}(\theta^*). \tag{6.41}$$

Proof. For any $\hat{\theta} \in \hat{\Theta}$, set $\Delta = \hat{\theta} - \theta^*$. We have

$$\hat{G}_{j,j}|\Delta_j| \leqslant \max_{j\neq k} |\hat{G}_{j,k}||\Delta|_1 + |\hat{G}\Delta|_\infty, \ \forall 1 \leqslant j \leqslant M.$$

Then, with probability at least $1 - 2M^{-1}$ $\forall 1 \leqslant j \leqslant M$

$$\left(1 - \frac{3}{2}Lr\right)|\Delta_j| \leqslant \left(\frac{1}{3\beta s} + \frac{3}{2}Lr\right) 24\frac{sr}{\zeta(s)^2} + 3r. \tag{6.42}$$

If Assumption 6.5, then Assumption 6.1 is satisfied with $\zeta(s) = \sqrt{1 - \frac{1}{\beta}}$ (cf. Lemma 6.6). Assumption 6.10, (6.40) and (6.42) yield

$$\left(1 - \frac{\beta - 1}{4s\beta}\right)|\Delta_j| \leqslant \left(\frac{1}{3\beta s} + \frac{\beta - 1}{4s\beta}\right)\frac{24sr\beta}{\beta - 1} + 3r, \ \forall j.$$

Since $s \geqslant 2$ and $\beta > 1$, this yields the first result. The second result on the thresholded estimators follows in an obvious way. □

6.7 Appendix

We recall here some well-known results of the theory of empirical processes.

Theorem 6.7. *[Bousquet's version of Talagrand's concentration inequality [11]] Let X_i be independent variables in $\mathcal{X}$ distributed according to P, and $\mathcal{F}$ be a set of functions from $\mathcal{X}$ to $\mathbb{R}$ such that $\mathbb{E}(f(X)) = 0$, $\|f\|_\infty \leqslant c$ and $\|f\|^2 \leqslant \sigma^2$ for any $f \in \mathcal{F}$. Let $Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)$. Then with probability $1 - e^{-x}$, it holds that*

$$Z \leqslant \mathbb{E}(Z) + \sqrt{2x(n\sigma^2 + 2c\mathbb{E}(Z))} + \frac{cx}{3}.$$

Theorem 6.8. *[Symmetrization theorem [101], p. 108] Let $X_1, \ldots, X_n$ be independent random variables with values in $\mathcal{X}$, and let $\epsilon_1, \ldots, \epsilon_n$ be a Rademacher sequence independent of $X_1, \ldots, X_n$. Let $\mathcal{F}$ ba a class of real-valued functions on $\mathcal{X}$. Then*

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n (f(X_i) - \mathbb{E}(f(X_i)))\right|\right) \leqslant 2\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \epsilon_i f(X_i)\right|\right).$$

Theorem 6.9. *[Contraction theorem [66], p. 95]. Let $x_1, \ldots, x_n$ be nonrandom elements of $\mathcal{X}$, and let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$. Consider Lipschitz functions $\gamma_i :\longrightarrow \mathbb{R}$, that is,*

$$|\gamma_i(s) - \gamma_i(s')| \leqslant |s - s'|, \quad \forall s, s' \in \mathbb{R}.$$

Let $\epsilon_1, \ldots, \epsilon_n$ be a Rademacher sequence. Then for any function $f^* : \mathcal{X} \to \mathbb{R}$, we have

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \epsilon_i(\gamma_i(f(x_i)) - \gamma_i(f^*(x_i)))\right|\right) \leqslant 2\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \epsilon_i((f(x_i) - f^*(x_i))\right|\right).$$

Chapter 7

Generalized Mirror Averaging and D-convex Aggregation

We study the problem of aggregation of estimators. Given a collection of M different estimators, we construct a new estimator, called aggregate, which is nearly as good as the best linear combination over a l_1-ball of $\mathbb{R}^M$ of the initial estimators. The aggregate is obtained by a particular version of the mirror averaging algorithm. We show that our aggregation procedure satisfies sharp oracle inequalities under general assumptions. Then we apply these results to a new aggregation problem: D-convex aggregation. Finally we implement our procedure in a Gaussian regression model with random design and we prove its optimality in a minimax sense up to a logarithmic factor. The results of Sections 7.1-7.6 are published in [70].

7.1 Introduction

We study the problem of aggregation in the framework introduced by Nemirovski [84], Catoni [17, 18] and Yang [109]. Given a collection of $M \geqslant 2$ different estimators, we would like to combine them to construct an improved estimator for a certain risk criterion. This new estimator is called aggregate. There exist three main aggregation problems: model selection (MS) aggregation, linear (L) aggregation and convex (C) aggregation. The objective of (MS) aggregation is to build an aggregate which is nearly as good as the best one among the M initial estimators ; that of (L) (respectively (C)) aggregation is to build an aggregate which is nearly as good as the best linear (respectively convex) combination of the M initial estimators. Consider the simplex $\Lambda^M = \{\lambda \in \mathbb{R}^M : \sum_{j=1}^M \lambda^{(j)} \leqslant 1, \lambda^{(j)} \geqslant 0, j = 1, \ldots, M\}$. Let Λ_D^M be the set of all $\lambda \in \Lambda^M$ such that λ has at most D nonzero components where D is an integer satisfying $1 \leqslant D \leqslant M$. In this chapter, we treat a new aggregation problem: D-convex aggregation which aims at constructing an aggregate which is nearly as good as the best convex combination of at most D of the M initial estimators. If $D = 1$, we recognize the (MS) aggregation problem ; if $D = M$, this is the (C) aggregation problem. The motivation of D-convex aggregation is twofold. First the approach is "sparse" if D is small and thus only a small number of estimators is selected among the M initial estimators. Second, as we will see it later, the rate of D-convex aggregation is faster than that of (C) aggregation. In a regression framework, Tsybakov [97] defines a notion of optimal rate of aggregation via a minimax approach and determines the optimal rates for (MS), (L) and (C) aggregation problems under strong assumptions on the design. This definition and the results of [97] will be recalled later. The literature on the aggregation of arbitrary estimators is quite extensive. We mention here only some recent works on the subject [5, 15, 68, 75, 92, 106, 110, 115].

Aggregation problems can be viewed as stochastic optimization problems, see [113, 57]. Let Z be a random variable with values in $\mathcal{Z}$, where $(\mathcal{Z}, \mathcal{A})$ is a measurable space. The unknown distribution of Z is denoted by P and the corresponding expectation by E. Let $\mathcal{D}_n = \{Z_1, \ldots, Z_n\}$ be an i.i.d sample of random variables Z_i taking their values in $\mathcal{Z}$ and distributed as Z. The distribution of $\mathcal{D}_n$ is denoted by P_n, the corresponding expectation by E_n. Let Λ be a bounded subset of $\mathbb{R}^M$. Consider a loss function Q: $\mathcal{Z} \times \Lambda \to \mathbb{R}^+$. The corresponding risk function $A : \Lambda \to \mathbb{R}^+$ is defined by $A(\lambda) = EQ(Z, \lambda)$, assuming that the expectation exists for all $\lambda \in \Lambda$. Direct minimization of A over $\lambda \in \Lambda$ is not possible since P is unknown. The problem of stochastic optimization consists in the construction of an estimator $\tilde{\lambda}_n$ measurable function of the sample $\mathcal{D}_n$ mimicking the oracle risk $\min_{\lambda \in \Lambda} A(\lambda)$,

i.e., such that

$$E_n A(\tilde{\lambda}_n) \leqslant \min_{\lambda \in \Lambda} A(\lambda) + \Delta_{n,M}^{\Lambda}, \tag{7.1}$$

where the remainder term $\Delta_{n,M}^{\Lambda} > 0$ should be as small as possible.

Mirror descent algorithms have been introduced by Nemirovski and Yudin [85]. In [113] a mirror averaging algorithm is implemented for the (C) aggregation problem and an oracle inequality of type (7.1) is obtained with a remainder term $\Delta_{n,M}^{\Lambda} \asymp \sqrt{\log(M)/n}$. The paper [57] implements a particular version of the mirror averaging algorithm for the (MS) aggregation problem and obtains an oracle inequality of type (7.1) with a remainder term $\Delta_{n,M}^{\Lambda} \asymp \log(M)/n$. These results are stronly related to previous work on the online prediction of individual deterministic sequences [20, 50, 60, 102] and also to the recent advances in the PAC-Bayesian theory [2, 18, 77].

In this chapter we consider a "continuous" version of the mirror averaging algorithm. This method for the squared loss has been studied in [17, 18] and [13] and recently for the general loss in [90]. We obtain oracle inequalities of type (7.1) over any l_1-ball of $\mathbb{R}^M$. As a consequence, we get an oracle inequality for the D-convex aggregation problem with the optimal rate of aggregation

$$\Delta_{n,M}^{\Lambda} \asymp \frac{D}{n}\left(\log\left(\frac{eM}{D}\right) + \log(n)\right). \tag{7.2}$$

The chapter is organized as follows. In Section 7.2, we present our "continuous" mirror averaging algorithm. In Section 7.3, we derive some preliminary results. In Section 7.4, we prove general oracle inequalities of type (7.1) satisfied by our procedure. In Section 7.5, we apply the results of the previous section to a regression model. In Section 7.6, we derive lower bounds on the optimal rate of D-convex aggregation for the considered regression model. Finally in Section 7.7, we investigate the use of different prior distribution to derive sparsity oracle inequalities of type:

$$E_n A(\tilde{\lambda}_n) \leqslant \min_{\lambda \in \Lambda}\left\{A(\lambda) + C\frac{M(\lambda)}{n}\log(Mn)\right\},$$

where $C > 0$ and for any $\lambda \in \mathbb{R}^M$, $M(\lambda)$ denotes the number of nonzero components of λ.

7.2 Generalized mirror averaging

Juditsky, Rigollet and Tsybakov [57] propose a mirror averaging algorithm working on a finite set. Here we propose a generalization of their mirror averaging algorithm that works on any bounded set $\Lambda \subset \mathbb{R}^M$.

Consider the following auxiliary stochastic optimization problem. Let $\mathcal{M}_1(\Lambda)$ be the set of all probability measures on a bounded subset Λ of $\mathbb{R}^M$. Consider $P' \in \mathcal{M}_1(\Lambda)$, denote by $\mathbb{E}_{P'}$ the corresponding expectation. Define the averaged loss $\overline{Q} : \mathcal{Z} \times \mathcal{M}_1(\Lambda) \to \mathbb{R}^+$ by

$$\overline{Q}(Z, P') = \mathbb{E}_{P'}(Q(Z, \lambda)) = \int_\Lambda Q(Z, \lambda) dP'(\lambda). \tag{7.3}$$

The corresponding averaged risk $\overline{A} : \mathcal{M}_1(\Lambda) \to \mathbb{R}^+$ is defined by

$$\overline{A}(P') = E(\overline{Q}(Z, P')). \tag{7.4}$$

Clearly

$$\overline{A}(P') = \mathbb{E}_{P'}(A(\lambda)) = \int_\Lambda A(\lambda) dP'(\lambda). \tag{7.5}$$

Direct minimization of $\overline{A}$ over $P' \in \mathcal{M}_1(\Lambda)$ is not possible since the distribution P of Z is unknown. The new stochastic optimization problem consists in minimization of $\overline{A}(P')$ over $P' \in \mathcal{M}_1(\Lambda)$, given the sample $\mathcal{D}_n$.

We now define our algorithm. Let $\mathcal{C}_b(\Lambda)$ be the space of real continuous bounded functions on Λ. For $\Pi \in \mathcal{M}_1(\Lambda)$, $\beta > 0$ and $\psi \in \mathcal{C}_b(\Lambda)$, consider the distribution $G_{\beta,\Pi}(\psi) \in \mathcal{M}_1(\Lambda)$ admitting the density $\frac{e^{-\psi/\beta}}{\int_\Lambda e^{-\psi/\beta} d\Pi}$ w.r.t. the distribution Π.

- Fix the initial distribution $\Pi \in \mathcal{M}_1(\Lambda)$ and take $\xi_0 = 0 \in \mathcal{C}_b(\Lambda)$.
- for $i = 1, \ldots, n$, do the recursive update

$$\xi_i(\cdot) = \xi_{i-1}(\cdot) + Q(Z_i, \cdot), \tag{7.6}$$

$$\overline{P}_i = G_{\beta,\Pi}(\xi_i). \tag{7.7}$$

- Compute the averaged distribution

$$\tilde{P}_n = \frac{1}{n} \sum_{i=1}^n \overline{P}_{i-1} \text{ (with } P_0 = \Pi). \tag{7.8}$$

- Compute $\tilde{\lambda}_n$

$$\tilde{\lambda}_n = \int_\Lambda \lambda d\tilde{P}_n(\lambda). \tag{7.9}$$

The procedure depends on two parameters, the "temperature" β and the prior distribution Π. It uses the sample $\mathcal{D}_n$ to update sequentially the prior distribution Π and computes the estimator $\tilde{\lambda}_n$ under the a posteriori distribution $\tilde{P}_n$. We will see later how the parameters are tuned depending on the problem to solve.

7.3 Preliminary results

We recall that the Kullback divergence between two distributions P and Q is defined by

$$K(P,Q) = \begin{cases} \int \log(\frac{dP}{dQ}) dP, & \text{if } P \ll Q \\ +\infty, & \text{elsewhere.} \end{cases}$$

Let $\lambda = (\lambda^{(1)}, \ldots, \lambda^{(M)})$ be a vector of $\mathbb{R}^M$ and denote the l_1-norm of λ by $||\lambda||_1 = \sum_{j=1}^M |\lambda^{(j)}|$. We denote by $B_1^M(0,c)$ the l_1-ball of $\mathbb{R}^M$ of radius c centered at 0. We make the following assumption on the risk function A

Assumption 7.1. ***(A)*** *There exists a constant $L_1 > 0$ such that $\forall \lambda, \lambda' \in \Lambda$,*

$$|A(\lambda) - A(\lambda')| \leqslant L_1 ||\lambda - \lambda'||_1.$$

Lemma 7.1. *Consider the subset $\Lambda = B_1^M(0,c)$ for a constant $c \geqslant 1$ and assume that (**A**) is satisfied on Λ. Then for all ϵ such that $0 < \epsilon \leqslant 1$, there exist a bounded subset $\Lambda_\epsilon \subset \mathbb{R}^M$ and a probability distribution P^* over Λ_ϵ such that*

$$\begin{cases} \Lambda \subset \Lambda_\epsilon, \\ K(P^*, \Pi) \leqslant M \log\left(\frac{1+\epsilon}{\epsilon}\right), \\ \overline{A}(P^*) \leqslant \min_{\lambda \in \Lambda} A(\lambda) + C'\epsilon, \end{cases} \tag{7.10}$$

where Π is the uniform distribution over Λ_ϵ and $0 < C' < \infty$ is a constant independent of M and ϵ.

Note that for any integer $d \geqslant 1$ and $a \geqslant 0$ the volume of the l_1 ball $B_1^d(0,a)$ in $\mathbb{R}^d$ is $C(d)a^d$ with

$$C(d) = \frac{\sqrt{d+1}}{d! 2^{\frac{d}{2}}}.$$

Proof. Take $\Lambda_\epsilon = B_1^M(0, c(1+\epsilon))$ and for Π the uniform probability distribution over Λ_ϵ. Then Π admits the following density

$$p(x) = \frac{1}{C(M)(c(1+\epsilon))^M} \mathbb{1}_{\Lambda_\epsilon}(x), \, \forall x \in \mathbb{R}^M,$$

w.r.t. the Lebesgue measure on $\mathbb{R}^M$. Note that $||p||_\infty \leqslant 1$ since $c \geqslant 1$. Let $\lambda^* \in \arg\min_{\lambda \in \Lambda} A(\lambda)$. Take for P^* the distribution admitting the following density

$$p^*(x) = \frac{1}{C(M)(c\epsilon)^M} \mathbb{1}_{B_1^M(\lambda^*, c\epsilon)}(x), \, \forall x \in \mathbb{R}^M,$$

w.r.t. the Lebesgue measure. We have then that

$$\begin{aligned} K(P^*, \Pi) &= \int_{\Lambda_\epsilon} p^*(\lambda) \log\left(\frac{p^*(\lambda)}{p(\lambda)}\right) d\lambda \\ &\leqslant M \log\left(\frac{1+\epsilon}{\epsilon}\right). \end{aligned}$$

Clearly

$$\min_{\lambda\in\Lambda} A(\lambda) \leqslant \overline{A}(P^*) = \int_{\Lambda_\epsilon} A(\lambda) p^*(\lambda) d\lambda.$$

Next

$$\begin{aligned} \overline{A}(P^*) - A(\lambda^*) &= \int_{\Lambda_\epsilon} A(\lambda) p^*(\lambda) d\lambda - A(\lambda^*) \\ &= \int_{\Lambda_\epsilon} (A(\lambda) - A(\lambda^*)) p^*(\lambda) d\lambda \\ &= \int_{B_1^M(\lambda^*, c\epsilon)} (A(\lambda) - A(\lambda^*)) p^*(\lambda) d\lambda \end{aligned}$$

because the support of p^* is $B_1^M(\lambda^*, c\epsilon)$. Assumption **(A)** yields

$$\begin{aligned} \overline{A}(P^*) - A(\lambda^*) &\leqslant L_1 \int_{B_1^M(\lambda^*, c\epsilon)} ||\lambda - \lambda^*||_1 p^*(\lambda) d\lambda \\ &\leqslant L_1 c\epsilon, \end{aligned}$$

which yields the result with $C' = L_1 c$. □

In most of the interesting problems, we have $M \gg n$. If we take the prior distribution Π as in Lemma 7.1, then we will obtain an oracle inequality of type (7.1) with a remainder term of the order M/n. We can improve this upper bound if we exploit the sparsity of the minimizer λ^*. Suppose the number of non zero components of λ^* , say D, satisfies $D \leqslant n$. If we know D, then we can estimate $A(\lambda^*)$ up to a remainder of the order $D \log(eM/D)/n$. For $1 \leqslant j \leqslant M$, denote by e_j the vector $(0, \ldots, 1, \ldots, 0) \in \mathbb{R}^M$ where 1 appears in jth position. Denote by $\mathcal{E}^M$ the set $\{e_1, \ldots, e_M\}$. Consider the set $\mathcal{I}$ of all subsets of $\mathcal{E}^M$ of cardinality D. This is a finite set of cardinality

$$|\mathcal{I}| = \binom{M}{D} \leqslant \left(\frac{eM}{D}\right)^D.$$

Denote $\mathcal{I} = \{I_1, \ldots, I_{|\mathcal{I}|}\}$. Every nonzero element $\lambda \in \Lambda_D^M$ can be associated with a pair $(I_j, \tilde{\lambda})$, where $I_j = \{e_{j_1}, \ldots, e_{j_D}\}$ is such that $\lambda^{(k)} = 0$ if k does not belong to I_j and $\tilde{\lambda} = (\tilde{\lambda}^{(1)}, \ldots, \tilde{\lambda}^{(D)})$ with $\tilde{\lambda}^{(i)} = \lambda^{(j_i)}$, $\forall 1 \leqslant i \leqslant D$. This representation is not unique but this is not important for further considerations. From now on, we take $\Lambda_D^M = \mathcal{I} \times \Lambda^D$, where $\Lambda^D = \{\lambda \in \mathbb{R}^D : \sum_{j=1}^D \lambda^{(j)} \leqslant 1, \lambda^{(j)} \geqslant 0, j = 1, \ldots, D\}$.

Lemma 7.2. *Assume that (**A**) holds on Λ_D^M. Then for all $\epsilon > 0$, there exist a bounded subset Λ_ϵ of $\mathbb{R}^M$ and a probability distribution P^* on Λ_ϵ such that*

$$\begin{cases} \Lambda_D^M \subset \Lambda_\epsilon, \\ K(P^*, \Pi) \leqslant D \log\left(\frac{eM}{D}\right) + D \log\left(\frac{1+\epsilon}{\epsilon}\right), \\ \overline{A}(P^*) \leqslant \min_{\lambda \in \Lambda_D^M} A(\lambda) + L_1 \epsilon, \end{cases} \tag{7.11}$$

where Π is the uniform distribution over Λ_ϵ.

Proof. Let $\lambda \in \mathbb{R}^D$, the l_1-norm of λ is defined by $||\lambda||_1 = \sum_{j=1}^D |\lambda^{(j)}|$. Let $\lambda^* = \arg\min_{\Lambda_D^M} A(\lambda)$. We note $\lambda^* = (I^*, \tilde{\lambda})$, where $I^* \in \mathcal{I}$ and $\tilde{\lambda} \in \Lambda^D$.

For $\epsilon > 0$, take $\Lambda_\epsilon = \mathcal{I} \times B_1^D(0, 1+\epsilon)$. Clearly $\Lambda_D^M \subset \Lambda_\epsilon$. The distribution Π is uniform over Λ_ϵ, so it admits the density

$$p(I_j, \lambda') = \frac{1}{|\mathcal{I}|} \mathbb{1}_{\mathcal{I}}(I_j) \frac{1}{C(D)((1+\epsilon))^D} \mathbb{1}_{B_1^D(0,1+\epsilon)}(\lambda'),\ \forall (I_j, \lambda') \in \mathcal{I} \times \mathbb{R}^D,$$

w.r.t. the measure product $\delta \times \nu$, where δ is the counting measure on $\mathcal{I}$ and ν the Lebesgue measure on $\mathbb{R}^D$.

Consider now the probability distribution P^* admitting the density

$$p^*(I_j, \lambda') = \mathbb{1}_{I^*}(I_j) \frac{1}{C(D)(\epsilon)^D} \mathbb{1}_{B_1^D(\tilde{\lambda}, c\epsilon)}(\lambda')\, \forall (I_j, \lambda') \in \mathcal{I} \times \mathbb{R}^D$$

w.r.t. the measure $\delta \times \nu$. Easy computation yields

$$K(P^*, \Pi) \leqslant \log(|\mathcal{I}|) + D \log\left(\frac{1+\epsilon}{\epsilon}\right).$$

Since $|\mathcal{I}| \leqslant \left(\frac{eM}{D}\right)^D$, we have

$$K(P^*, \Pi) \leqslant D \log\left(\frac{eM}{D}\right) + D \log\left(\frac{1+\epsilon}{\epsilon}\right).$$

Under Assumption (**A**), following the proof of Lemma 7.1 yields

$$0 \leqslant \overline{A}(P^*) - \min_{\lambda \in \Lambda_D^M} A(\lambda) \leqslant L_1 \epsilon.$$

□

7.4 General oracle risk inequalities

For $\Pi \in \mathcal{M}_1(\Lambda)$ and $\beta > 0$, define the functional $W_{\beta,\Pi} : \mathcal{C}_b(\Lambda) \to \mathbb{R}$ by

$$W_{\beta,\Pi}(\psi) = \beta \log(\mathbb{E}_\Pi[e^{-\psi/\beta}]) = \beta \log\left(\int_\Lambda e^{-\psi(\lambda)/\beta} d\Pi(\lambda)\right),\ \forall \psi \in \mathcal{C}_b(\Lambda). \tag{7.12}$$

For all $P' \in \mathcal{M}_1(\Lambda)$ and $\psi \in \mathcal{C}_b(\Lambda)$,

$$\beta K(P', \Pi) + W_{\beta,\Pi}(\psi) \geqslant -\int_\Lambda \psi dP'. \tag{7.13}$$

This result can be found in [27] page 264. This is a direct application of the Jensen's inequality. If P' is absolutely continuous w.r.t. Π, denote $p = dP'/d\Pi$,

$$\begin{aligned} \log\left(\int_\Lambda e^{-\psi(\lambda)/\beta} d\Pi(\lambda)\right) &\geqslant \int_\Lambda \log\left(\frac{1}{p} e^{-\psi(\lambda)/\beta} dP'(\lambda)\right) \\ &= -K(P', \Pi) - \frac{1}{\beta}\int_\Lambda \psi dP'. \end{aligned}$$

If P' is not absolutely continuous w.r.t. Π, the inequality is trivial since $K(P', \Pi) = \infty$.

Consider the function Q_1 defined on $\mathcal{Z} \times \Lambda \times \Lambda$ by $Q_1(z, \lambda, \lambda') = Q(z, \lambda) - Q(z, \lambda')$ for all $z \in \mathcal{Z}$ and all $\lambda, \lambda' \in \Lambda$.

Theorem 7.1. *Consider $\Lambda = B_1^M(0, c)$ for a constant $c \geqslant 1$ and a loss function Q such that the associated risk A satisfies Assumption (**A**). Then for all $0 < \epsilon \leqslant 1$, the aggregate $\tilde{\lambda}_n$ obtained by the implementation of the mirror averaging algorithm over Λ_ϵ with the prior distribution Π, where Λ_ϵ and Π are defined in Lemma 7.1, satisfies*

$$E_{n-1} A(\tilde{\lambda}_n) \leqslant \min_{\lambda \in \Lambda} A(\lambda) + C'\epsilon + \frac{\beta M}{n} \log\left(\frac{1+\epsilon}{\epsilon}\right) + S_n, \tag{7.14}$$

where

$$S_n \triangleq \beta E_n \log\left(\int_{\Lambda_\epsilon} \exp\left[-\frac{Q_1(Z_n, \lambda, \tilde{\lambda}_n)}{\beta}\right] \tilde{P}_n(d\lambda)\right).$$

Proof. By definition of $W_{\beta,\Pi}(.)$, for $i = 1, \ldots, n$,

$$\begin{aligned} W_{\beta,\Pi}(\xi_i) - W_{\beta,\Pi}(\xi_{i-1}) &= \beta \log\left(\frac{\int_{\Lambda_\epsilon} e^{-\xi_i(\lambda)/\beta} d\Pi(\lambda)}{\int_{\Lambda_\epsilon} e^{-\xi_{i-1}(\lambda)/\beta} d\Pi(\lambda)}\right) \\ &= \beta \log\left(\int_{\Lambda_\epsilon} e^{-Q(Z_i,\lambda)/\beta} d\overline{P}_{i-1}(\lambda)\right). \end{aligned} \tag{7.15}$$

Taking the expectation on both sides of (7.15), summing up over i, using the fact that $(\overline{P}_{i-1}, Z_i)$ has the same distribution as $(\overline{P}_{i-1}, Z_n)$ for $i = 1, \ldots, n$ and applying the concavity of the logarithmic function, we get

$$\begin{aligned} \frac{E_n[W_{\beta,\Pi}(\xi_n) - W_{\beta,\Pi}(\xi_0)]}{n} &= E_n\left[\frac{\beta}{n}\sum_{i=1}^n \log\left(\int_{\Lambda_\epsilon} e^{-Q(Z_n,\lambda)/\beta} d\overline{P}_{i-1}(\lambda)\right)\right] \\ &\leqslant \beta E_n \log\left(\int_{\Lambda_\epsilon} e^{-Q(Z_n,\lambda)/\beta} d\tilde{P}_n(\lambda)\right) \triangleq S. \end{aligned} \tag{7.16}$$

Since $Q_1(z,\omega,\mathbb{E}_{\tilde{P}_n}[\omega]) = Q(z,\omega) - Q(z,\mathbb{E}_{\tilde{P}_n}[\omega])$ and $\mathbb{E}_{\tilde{P}_n}[\omega] = \tilde{\lambda}_n$, S satisfies

$$\begin{aligned} S &= \beta E_n \log\left(\mathbb{E}_{\tilde{P}_n} \exp\left[-\frac{Q(Z_n,\omega)}{\beta}\right]\right) \\ &= \beta E_n \log\left(\exp\left[-\frac{Q(Z_n,\mathbb{E}_{\tilde{P}_n}[\omega])}{\beta}\right]\right) + \beta E_n \log\left(\mathbb{E}_{\tilde{P}_n} \exp\left[-\frac{Q_1(Z_n,\omega,\mathbb{E}_{\tilde{P}_n}[\omega])}{\beta}\right]\right) \\ &= -E_{n-1}A(\tilde{\lambda}_n) + S_n. \end{aligned} \tag{7.17}$$

By (7.13), we have

$$\begin{aligned} -W_{\beta,\Pi}(\xi_n) &\leqslant \int_{\Lambda_\epsilon} \xi_n dP^* + \beta K(P^*,\Pi) \\ \frac{E_n[-W_{\beta,\Pi}(\xi_n)]}{n} &\leqslant E_n \int_{\Lambda_\epsilon} \frac{\xi_n}{n} dP^* + \frac{\beta}{n} K(P^*,\Pi) \\ \frac{E_n[-W_{\beta,\Pi}(\xi_n)]}{n} &\leqslant \overline{A}(P^*) + \frac{\beta}{n} K(P^*,\Pi). \end{aligned} \tag{7.18}$$

We used Fubini-Tonelli Theorem on the right-hand side at the last line. Combining Lemma 7.1, (7.16), (7.17) and (7.18) gives the result. □

If Λ is a finite set, we can suppress the term $\frac{\beta M}{n}\log(\frac{1+\epsilon}{\epsilon})$ in (7.1) and we get an inequality as in [57]. Similar result in a different context has been proved in [90].

Theorem 7.2. *For all $\epsilon > 0$ and any loss function Q such that the associated risk A satisfies Assumption (**A**), the aggregate $\tilde{\lambda}_n$ obtained by the implementation of the mirror averaging algorithm on the set Λ_ϵ with the prior distribution Π, where Λ_ϵ and Π are defined in Lemma 7.2, satisfies the inequality*

$$E_{n-1}A(\tilde{\lambda}_n) \leqslant \min_{\lambda\in\Lambda_D^M} A(\lambda) + L_1\epsilon + \beta\frac{D}{n}\left(\log\left(\frac{1+\epsilon}{\epsilon}\right) + \log\left(\frac{eM}{D}\right)\right) + S_n, \tag{7.19}$$

where

$$S_n \triangleq \beta E_n \log\left(\int_{\Lambda_\epsilon} \exp\left[-\frac{Q_1(Z_n,\lambda,\tilde{\lambda}_n)}{\beta}\right]\tilde{P}_n(d\lambda)\right).$$

Proof. The proof is exactly the same as that of Theorem 7.1 until the inequality (7.18). It suffices then to apply Lemma 7.2, (7.16), (7.17) and (7.18) to get the result. □

7.5 Oracle inequalities for Gaussian regression with random design

Let $\mathcal{X}$ be a Borel subset of $\mathbb{R}^d$, (X,Y) a random vector such that $X \in \mathcal{X}$, $Y \in \mathbb{R}$ and

$$Y = f(X) + \xi, \tag{7.20}$$

where the regression function f is unknown. Suppose that the error ξ is a gaussian random variable $\mathcal{N}(0,\sigma^2)$ independent of X. Let P_f (respectively P^X) be the distribution of (X,Y) (respectively X). Suppose $||f||_\infty \leqslant L$ for a finite constant L where $||f||_\infty = \inf\{C : |f(x)| \leqslant C \text{ a.s. on } \mathcal{X}\}$. Let $||f|| = \left(\int_{\mathcal{X}} f^2(x)dP^X(x)\right)^{1/2}$. Let $\hat{f}_n$ be an estimator of f built from the i.i.d sample $\{(X_1,Y_1),\dots,(X_n,Y_n)\}$, where (X_i,Y_i) is distributed as (X,Y). Let P_f^n be the distribution of $\{(X_1,Y_1),\dots,(X_n,Y_n)\}$ and E_f^n the corresponding expectation. The risk L_2 of $\hat{f}_n$ is $E_f^n||\hat{f}_n - f||^2$.

Consider M Borel functions $f_1,\dots,f_M : \mathcal{X} \to \mathbb{R}$ satisfying $||f_j||_\infty \leqslant L$, $1, \leqslant j \leqslant M$. Let Λ be a Borel subset of $\mathbb{R}^M$, let $\lambda = (\lambda^{(1)},\dots,\lambda^{(M)}) \in \Lambda$, we denote by f_λ the linear combination $\sum_{j=1}^M \lambda^{(j)} f_j$. We recall the definition of the optimal rate of aggregation given in [97].

Definition 7.1. *Consider the gaussian regression model with random design (7.20), let Λ be a bounded subset of $\mathbb{R}^M$ and $\mathcal{F}_0$ the space of Borel functions $g : \mathcal{X} \to \mathbb{R}$ bounded by a finite constant L. A positive sequence of numbers $\psi_{n,M}^\Lambda$ is called optimal rate of aggregation for $(\Lambda,\mathcal{F}_0)$ if*

- *for all distributions P^X and any set of functions $\{f_\lambda,\ \lambda\in\Lambda\}$ indexed by Λ and contained in $\mathcal{F}_0$, there exists an estimator $\tilde{f}_n$ of f (aggregate) such that*

$$\sup_{f\in\mathcal{F}_0}[E_f^n||\tilde{f}_n - f||^2 - \inf_{\lambda\in\Lambda}||f_\lambda - f||^2] \leqslant C\psi_{n,M}^\Lambda,\ \forall n \geqslant 1, \tag{7.21}$$

  for a constant $C < \infty$ independent of n and M.

- *There exists a distribution P^X and a set of functions $\{f_\lambda,\ \lambda\in\Lambda\}$ indexed by Λ and contained in $\mathcal{F}_0$ such that for all estimators T_n of f,*

$$\sup_{f\in\mathcal{F}_0}[E_f^n||T_n - f||^2 - \inf_{\lambda\in\Lambda}||f_\lambda - f||^2] \geqslant c\psi_{n,M}^\Lambda,\ \forall n \geqslant 1, \tag{7.22}$$

  for a constant $c > 0$ independent of n and M.

Tsybakov [97] determines the optimal rate of aggregation $\psi_{n,M}^\Lambda$ for the (L), (C), (MS) aggregation problems under some strong assumptions on P^X

$$\psi_{n,M}^\Lambda \asymp \begin{cases} \frac{M}{n} & \text{for (L) aggregation,} \\ \frac{M}{n} & \text{for (C) aggregation, if } M \leqslant \sqrt{n}, \\ \sqrt{\{\log(eM/\sqrt{n})\}/n} & \text{for (C) aggregation, if } M > \sqrt{n}, \\ \frac{\log(M)}{n} & \text{for (MS) aggregation.} \end{cases} \tag{7.23}$$

Although these rates do not satisfy Definition 7.1 since they were not derived for all distributions P^X, they serve as a benchmark for more general problems. In this chapter, the

aggregation procedures we propose work for all distributions P^X. Furthermore we prove they are optimal up to a logarithmic factor in the sense of Definition 7.1.

We now apply Theorems 7.1 and 7.2 to treat (C) and D-convex aggregation with the L_2 loss in the regression case. Take $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ and $Z = (X, Y)$ with $X \in \mathcal{X}$ and $Y \in \mathbb{R}$ satisfying (7.20). Here $P = P_f$ and $E = E_f$. Define the loss function $Q : \mathcal{Z} \times \Lambda \to \mathbb{R}^+$ by

$$Q(z, \lambda) = (y - f_\lambda(x))^2, \forall \lambda \in \Lambda,$$

where $z = (x, y) \in \mathcal{X} \times \mathbb{R}$ and $H(x) = (f_1(x), \ldots, f_M(x))^T$. To apply theorems 7.1 and 7.2, we need first to check assumption (**A**).

Lemma 7.3. *The risk A satisfies Assumption (**A**) on $\Lambda = B_1^M(0, c)$ with constant $L_1 = 2L^2(1 + c)$.*

Proof. The risk A satisfies

$$\begin{aligned} A(\lambda) - A(\lambda') &= E[(Y - f_\lambda(X))^2 - (Y - f_{\lambda'}(X))^2] \\ &= E[(2Y - (f_\lambda(X) + f_{\lambda'}(X)))\,(f_\lambda(X) - f_{\lambda'}(X))] \\ &= E[(2f(X) - (f_\lambda(X) + f_{\lambda'}(X)))\,(f_\lambda(X) - f_{\lambda'}(X))] \\ |A(\lambda) - A(\lambda')| &\leqslant E[(2|f(X)| + L(||\lambda||_1 + ||\lambda'||_1))]L||\lambda - \lambda'||_1. \end{aligned}$$

Recall $||f||_\infty \leqslant L$ and $||\lambda||_1 \leqslant c$, $\forall \lambda \in \Lambda$. Thus,

$$|A(\lambda) - A(\lambda')| \leqslant 2L^2(1 + c)||\lambda - \lambda'||_1.$$

□

The following lemma is a result of [57]. For the sake of completeness, we reproduce the proof in the appendix up to some minor modifications to fit our model.

Lemma 7.4. *For the regression model (7.20) and $\Lambda = B_1^M(0, c)$, we have $S_n \leqslant 0$, $\forall \beta \geqslant 2(\sigma^2 + ((c + 1)L)^2)$.*

Now we define our convex aggregate $\tilde{f}_n^C$. We consider two cases $M \leqslant \sqrt{n}$ and $M > \sqrt{n}$. In the case $M \leqslant \sqrt{n}$, then $\tilde{f}_n^C = f_{\tilde{\lambda}_n}$ where $\tilde{\lambda}_n$ is defined in Theorem 7.1. Note that in this case, $\tilde{\lambda}_n$ is built using the continuous mirror averaging algorithm introduced in Section 7.2. In the case $M > \sqrt{n}$, let m be defined by

$$m = \left\lceil \sqrt{n} / \sqrt{\log(eM/\sqrt{n})} \right\rceil. \tag{7.24}$$

Clearly $1 \leqslant m \leqslant M$. Denote by $\mathcal{G}$ the finite set of functions h of the form

$$h(x) = \frac{1}{m} \sum_{j=1}^{M} k_j f_j(x),\ k_j \in \{0, \ldots, m\},\ \sum_{j=1}^{M} k_j \leqslant m.$$

We have

$$|\mathcal{G}| = \sum_{j=1}^{m} \binom{M+j-1}{j} \leqslant \left(e \frac{M+m-1}{m} \right)^m,$$

see for instance [39] page 38. Consider the (MS) aggregation problem over $\mathcal{G}$. Let $\hat{\lambda}_n$ be the discrete mirror averaging estimator introduced in [57]. Then for $M > \sqrt{n}$, $\tilde{f}_n^C = f_{\hat{\lambda}_n}$. So the convex aggregate is given by

$$\tilde{f}_n^C(x) = \begin{cases} f_{\tilde{\lambda}_n}(x) & \text{if } M \leqslant \sqrt{n}, \\ f_{\hat{\lambda}_n}(x) & \text{if } M > \sqrt{n}. \end{cases} \tag{7.25}$$

Theorem 7.3. *For any distribution P^X, the aggregate $\tilde{f}_n^C$ satisfies the following oracle inequality*

$$E_f^n ||\tilde{f}_n^C - f||^2 \leqslant \min_{\lambda \in \Lambda^M} ||f_\lambda - f||^2 + C(\sigma, L) \Delta_{n,M}^C, \tag{7.26}$$

where $C(\sigma, L)$ is a constant depending only on σ and L and

$$\Delta_{n,M}^C = \begin{cases} \frac{M}{n} \log(n) & \text{if } M \leqslant \sqrt{n}, \\ \sqrt{\{\log(eM/\sqrt{n})\}/n} & \text{if } M > \sqrt{n}. \end{cases}$$

$\Delta_{n,M}^C$ is the optimal rate of (C) aggregation up to the logarithmic factor $\log(n)$.

Proof. If $M \leqslant \sqrt{n}$, take $\epsilon = 1/n$ and apply Theorem 7.1 along with Lemmas 7.3 and 7.4. If $M > \sqrt{n}$, the following approximation result can be obtained by the Maurey argument (see [84], pages 192,193)

$$\min_{g \in \mathcal{G}} ||g - f||^2 \leqslant \min_{\lambda \in \Lambda^M} ||f_\lambda - f||^2 + \frac{L^2}{m}, \tag{7.27}$$

the above formula means that the minimum over the finite set $\mathcal{G}$ approximates the minimum over the convex combinations of the functions $f_1, \ldots, f_M$ up to the term L^2/m. Apply Corollary 4.5 of [57] along with (7.27) to get the result. For the lower bounds, see [97]. □

We define now the D-convex aggregate $\tilde{f}_n^D$. We recall that the D-convex aggregate is motivated by the fact that M can be much larger than n in some applications. So we suppose now that $M \gg n$. We consider two cases $D \leqslant \sqrt{n/\log(eM/\sqrt{n})}$ and $D > \sqrt{n/\log(eM/\sqrt{n})}$. In the first case, $\tilde{f}_n^D = f_{\tilde{\lambda}_n}$ where $\tilde{\lambda}_n$ is defined in Theorem 7.2. In the

second case, $\tilde{f}_n^D = f_{\hat{\lambda}_n}$, where we use the same $\hat{\lambda}_n$ as the one defined for the convex aggregate above. So, the D-convex aggregate $\tilde{f}_n^D$ is defined by

$$\tilde{f}_n^D(x) = \begin{cases} f_{\tilde{\lambda}_n}(x) & \text{if } D \leqslant \sqrt{\frac{n}{\log(eM/\sqrt{n})}}, \\ f_{\hat{\lambda}_n}(x) & \text{if } D > \sqrt{\frac{n}{\log(eM/\sqrt{n})}}. \end{cases} \tag{7.28}$$

Theorem 7.4. *For any distribution P^X, the aggregate $\tilde{f}_n^D$ satisfies the following oracle inequality*

$$E_f^n ||\tilde{f}_n^D - f||^2 \leqslant \min_{\lambda \in \Lambda_D^M} ||f_\lambda - f||^2 + C(\sigma, L)\Delta_{n,M,D}, \tag{7.29}$$

where $C(\sigma, L)$ is a constant depending only on σ and L and

$$\Delta_{n,M,D} = \begin{cases} \frac{D}{n}\left(\log\left(\frac{eM}{D}\right) + \log(n)\right) & \text{if } D \leqslant \sqrt{\frac{n}{\log(eM/\sqrt{n})}}, \\ \sqrt{\{\log(eM/\sqrt{n})\}/n} & \text{if } D > \sqrt{\frac{n}{\log(eM/\sqrt{n})}}. \end{cases}$$

If $M \geqslant n$, then $\Delta_{n,M,D}$ is the optimal rate of D-convex aggregation.

Proof. If $D \leqslant \sqrt{\frac{n}{\log(eM/\sqrt{n})}}$, take $\epsilon = 1/n$ and apply Theorem 7.2 along with Lemmas 7.3 and 7.4. If $D > \sqrt{\frac{n}{\log(eM/\sqrt{n})}}$, then $M > \sqrt{n}$. The result follows from Theorem 7.3 and the inclusion $\Lambda_D^M \subset \Lambda^M$. The lower bounds are derived in the next section. □

Like for (C) aggregation, we observe an elbow effect. For $D \leqslant \sqrt{\frac{n}{\log(eM/\sqrt{n})}}$ the rate of D-convex aggregation is better than the one of (C) aggregation. For $D > \sqrt{\frac{n}{\log(eM/\sqrt{n})}}$ the rates of D-convex and (C) aggregation are equal.

7.6 Lower bounds for D-convex aggregation in Gaussian regression model with random design

Now we derive lower bounds for D-convex aggregation in the regression model (7.20).

Theorem 7.5. *Consider the Gaussian regression model with random design (7.20), the optimal rate of D-convex aggregation denoted $\psi_{n,M,D}$ satisfies the following inequalities*

$$\psi_{n,M,D} \geqslant \begin{cases} c\frac{D}{n}\log\left(\frac{eM}{D}\right) & \text{if } D \leqslant \sqrt{\frac{n}{\log(eM/\sqrt{n})}}, \\ c\sqrt{\{\log(eM/\sqrt{n})\}/n} & \text{if } D > \sqrt{\frac{n}{\log(eM/\sqrt{n})}}, \end{cases} \tag{7.30}$$

for a constant $c > 0$ independent of n, M and D.

Thus in the regression setup (7.20), the rate of D-convex aggregation $\Delta_{n,M,D}$ in Theorem 7.4 is optimal if we suppose that $M \gg n$. The proof of Theorem 7.5 uses Lemma 4 of [9] and Lemma 1 of [97].

Proof. Consider a cube $\mathcal{S} \subset \mathcal{X}$. Take for P^X the uniform distribution over $\mathcal{S}$, then $L_2(\mathcal{X}, P^X) = L_2(S, dx)$. Let $(\phi_j)_{j=1,\dots,M}$ be an orthogonal set of functions of $L_2(S, dx)$ such that $||\phi_j||_\infty \leqslant A < \infty$, $\forall 1 \leqslant j \leqslant M$. Take $\phi_j(x) = A\cos(ajx_1 + b)$, for $x \in S$ and for proper constants A, a, b, where x_1 is the first coordinate of x. Define the functions $f_j(x) = \gamma\phi_j(x)1_S(x)$, $1 \leqslant j \leqslant M$, where γ is chosen small enough so that $||f_j||_\infty \leqslant L$, $\forall 1 \leqslant j \leqslant M$.

If $D \leqslant \sqrt{n/\log(eM/\sqrt{n})}$, consider the finite set

$$\mathcal{C}' = \left\{ \frac{1}{\sqrt{3}} \sqrt{\frac{\log(eM/D)}{n}} \sum_{j=1}^{M} \omega_j f_j \,:\, \omega \in \{0,1\}^M \text{ and } \sum_{j=1}^{M} \omega_j \leqslant D \right\}. \tag{7.31}$$

We check that $\mathcal{C}' \subset \{f_\lambda,\, \lambda \in \Lambda_D^M\}$. The function $x \to x\sqrt{\log eM/x}$ is nondecreasing on $[1, \infty)$. Next, for $D \leqslant \sqrt{\frac{n}{\log(eM/\sqrt{n})}}$ we have

$$\frac{D}{\sqrt{3}} \sqrt{\frac{\log(eM/D)}{n}} \leqslant \frac{1}{\sqrt{3}} \sqrt{\frac{n}{\log(eM/\sqrt{n})}} \frac{1}{\sqrt{n}} \sqrt{\log\left(\frac{eM}{\sqrt{n}} \sqrt{\log\left(\frac{eM}{\sqrt{n}}\right)}\right)}. \tag{7.32}$$

Using that for all $y > 0$, we have $\log(y\sqrt{\log(y)}) \leqslant 3\log(y)$, we obtain

$$\frac{D}{\sqrt{3}} \sqrt{\frac{\log(eM/D)}{n}} \leqslant 1.$$

Let f_λ and $f_{\lambda'}$ be in $\mathcal{C}'$. Then

$$||f_\lambda - f_{\lambda'}||^2 = \frac{\gamma^2}{3} \frac{\rho(\lambda, \lambda')}{n} \log(eM/D), \tag{7.33}$$

with $\rho(\lambda, \lambda') \leqslant 2D$. If $M \geqslant 6D$, Lemma 4 of [9] guarantees the existence of a subset $\mathcal{N} \subset \mathcal{C}'$ such that

$$\begin{cases} \log(|\mathcal{N}|) \geqslant \tilde{c} D \log(\frac{eM}{D}), \\ \rho(\lambda, \lambda') \geqslant \tilde{c} D,\ \forall \lambda,\ \lambda' \in \mathcal{N}. \end{cases} \tag{7.34}$$

For all λ and λ'

$$K(P_{f_\lambda}^n, P_{f_{\lambda'}}^n) = \frac{n}{2\sigma^2} ||f_\lambda - f_{\lambda'}||^2, \tag{7.35}$$

thus

$$K(P_{f_\lambda}^n, P_{f_{\lambda'}}^n) \leqslant c\gamma^2 \log(|\mathcal{N}|).$$

Take γ small enough so that $\mathcal{C}' \subset \mathcal{F}_0$. Then $\min_{\lambda\in\Lambda^M} ||f_\lambda - f||^2 = 0$ for all $f \in \mathcal{C}'$. Lemma 1 of [97] yields

$$\sup_{f\in\mathcal{C}'} E_f^n ||T_n - f||^2 \geqslant c\frac{D}{n}\log(eM/D).$$

If $D \leqslant M \leqslant 6D$, we have a lower bound $c\frac{M}{n}$ for a constant $c > 0$ independent of M, see [97]. Then it remains to note that for these values of M $\frac{M}{n} \asymp \frac{D}{n}\log(\frac{eM}{D})$.

If $D > \sqrt{n/\log(eM/\sqrt{n})}$, define an integer m by

$$m = \left\lceil c_2\sqrt{\frac{n}{\log(\frac{eM}{\sqrt{n}})}}\right\rceil,$$

where the constant c_2 is chosen small enough so that $M \geqslant 6m$ and $m \leqslant D$. Consider the finite set $\mathcal{C}''$ of convex combinations of $f_1, \dots, f_M$ such that m coefficients equal $1/m$ and the remaining $M - m$ equal zero. Clearly $\mathcal{C}'' \subset \{f_\lambda,\, \lambda \in \Lambda_D^M\}$.

Let f_λ and $f_{\lambda'}$ be in $\mathcal{C}''$, $||f_\lambda - f_{\lambda'}||^2 \leqslant 2\frac{\gamma^2}{m}$. Take γ small enough so that $\mathcal{C}'' \subset \mathcal{F}_0$ and $\min_{\lambda\in\Lambda^M} ||f_\lambda - f||^2 = 0$ for all $f \in \mathcal{C}''$. We then have to bound from below $\sup_{f\in\mathcal{C}''} E_f^n||T_n - f||^2$ by $c\sqrt{\frac{1}{n}\log(\frac{eM}{\sqrt{n}})}$. Lemma 4 of [9] guarantees the existence of a finite subset $\mathcal{N}$ of $\mathcal{C}''$ such that

$$\log(|\mathcal{N}|) \geqslant cm\log\left(\frac{eM}{m}\right), \tag{7.36}$$

for a constant $c > 0$ and such that for all functions $g_1, g_2 \in \mathcal{N}$,

$$||g_1 - g_2||^2 \geqslant c\gamma^2/m, \tag{7.37}$$

where $c > 0$ is a different constant. In view of (7.35), we have $K(P_{g_1}^n, P_{g_2}^n) \leqslant c\gamma^2 n/m \leqslant c\gamma^2\log(|\mathcal{N}|)$. Take γ small enough such that $c\gamma^2 < 1/16$ and apply Lemma 1 of [97]. □

7.7 Sparsity oracle inequality and choice of the prior Π

In Theorem 7.4, we derived an oracle inequality of the form

$$E_n A(\tilde{\lambda}_n) \leqslant \min_{\lambda\in\Lambda} A(\lambda) + C(L,c)\frac{D}{n}\log\left(\frac{eM}{D}\right)\log(n),$$

for the mirror averaging estimator in a Gaussian regression model. A drawback of this procedure is that the choice of the prior Π requires that the number of nonzero components of the target vector λ^* satisfies $M(\lambda^*) \leqslant D$ with **known** D. In this section, we want to build an estimator $\tilde{\lambda}_n$ satisfying the following sparsity oracle inequality

$$E_n A(\tilde{\lambda}_n) \leqslant \min_{\lambda\in\Lambda}\left\{A(\lambda) + C(L,c)\frac{M(\lambda)}{n}\log(Mn)\right\}, \tag{7.38}$$

where $\Lambda = B_1^M(0,c)$ and $M(\lambda) = |\{j \,:\, \lambda_j \neq 0\}|$. Dalalyan and Tsybakov [25] proved a sparsity oracle inequality of this type in a fixed design regression model when the set of parameters is a l_2 ball of $\mathbb{R}^M$ for the exponential weight aggregate. Dalalyan and Tsybakov [24] derive sparsity oracle inequalities for the mirror averaging in the same stochastic optimization problem with a Cauchy type choice for the prior distribution Π. They also provide practical rules for the tuning of the parameters such as the temperature β and the radius c of the simplex Λ. Here we derive the sparsity oracle inequality (7.38) in a general stochastic framework for the mirror averaging algorithm as in [24]. The difference from [24] is that we propose two other ways to choose the prior distribution Π. The first way is to take Π as a mixture of probability distributions on Λ instead of the uniform distribution on a Λ considered in Section 7.4. The second way is to put a prior distribution on the number D of nonzero components with more weight on the small values of D. The latter choice for Π is related to penalized model-selection techniques where models with a large number of parameters are more penalized than models with a small number of parameters.

7.7.1 Taking Π as a mixture of probability distributions

Fix $\epsilon > 0$. Recall that $\Lambda_\epsilon = B_1^M(0, c(1+\epsilon))$. Define the function $\pi_0 : \mathbb{R} \to \mathbb{R}$ by

$$\pi_0(t) = (1-\gamma)\frac{1}{2c\eta\epsilon}\mathbb{1}_{[-c\eta\epsilon, c\eta\epsilon]}(t) + \gamma\frac{1}{2c(1+\epsilon)}\mathbb{1}_{[-c(1+\epsilon), c(1+\epsilon)]}(t), \quad \forall t \in \mathbb{R},$$

where $\eta, \gamma > 0$ are some parameters to be tuned. We consider the prior distribution Π admitting the following density w.r.t. the Lebesgue measure on $\mathbb{R}^M$:

$$\pi(x) = \frac{1}{C_N}\prod_{j=1}^{M}\pi_0(x_j)\mathbb{1}_{\Lambda_\epsilon}(x), \; x = (x_1, \ldots, x_M) \in \mathbb{R}^M \tag{7.39}$$

where C_N is the normalizing constant. We have the following result.

Lemma 7.5. *Assume that $M \geqslant 2$. Take $\eta = \gamma = 1/M$. Consider the sets $\Lambda = B_1^M(0,c)$ and $\Lambda_\epsilon = B_1^M(0, c(1+\epsilon))$ for a constant $c \geqslant 1$ and the prior distribution Π defined in (7.39). Let Assumption (**A**) be satisfied on Λ. Then, for any $\lambda^* \in \Lambda$ and any ϵ such that $0 < \epsilon < 1$, there exists a probability distribution P^* supported on Λ_ϵ such that*

$$\begin{aligned} K(P^*, \Pi) &\leqslant M(\lambda^*)\left(2(\log M) + \log\left(\frac{1+\epsilon}{\epsilon}\right)\right) + 2\left(1 - \frac{M(\lambda^*)}{M}\right), \\ |\overline{A}(P^*) - A(\lambda^*)| &\leqslant 2L_1 c\epsilon. \end{aligned}$$

Proof. Let P^* be equal to the probability distribution $P_{\lambda^*,c\eta\epsilon}$ admitting the following density w.r.t. the Lebesgue measure:

$$p_{\lambda^*}(x) = \frac{1}{(2c\eta\epsilon)^M} \prod_{j=1}^{M} \mathbb{1}_{[\lambda_j^*-c\eta\epsilon,\lambda_j^*+c\eta\epsilon]}(x_j), \quad \forall x \in \mathbb{R}^M.$$

Since $\lambda^* \in \Lambda$ and $\eta = 1/M$, we have $P^* << \Pi$. Next,

$$\left| \int_\Lambda A(\lambda) p_{\lambda^*}(\lambda) d\lambda - A(\lambda^*) \right| \leqslant 2L_1 cM\eta\epsilon = 2L_1 c\epsilon.$$

For any j set $p_{\lambda_j^*}(x_j) = \frac{1}{2c\eta\epsilon} \mathbb{1}_{[\lambda_j^*-c\eta\epsilon,\lambda_j^*+c\eta\epsilon]}(x_j)$. We have

$$\begin{aligned} K(P^*,\Pi) &= \int_{\Lambda_\epsilon} p_{\lambda^*}(\lambda) \log\left(\frac{p_{\lambda^*}(\lambda)}{\pi(\lambda)}\right) d\lambda \\ &= \int_{B_\infty^M(\lambda^*,c\eta\epsilon)} p_{\lambda^*}(\lambda) \log\left(\frac{p_{\lambda^*}(\lambda)}{\pi(\lambda)}\right) d\lambda \\ &= \sum_{j=1}^{M} \int_{\lambda_j^*-c\eta\epsilon}^{\lambda_j^*+c\eta\epsilon} p_{\lambda_j^*}(\lambda_j) \log\left(\frac{p_{\lambda_j^*}(\lambda_j)}{\pi_0(\lambda_j)}\right) d\lambda_j + \int_{B_\infty^M(\lambda^*,c\eta\epsilon)} p_{\lambda^*}(\lambda) \log(C_N) d\lambda, \quad (7.40) \end{aligned}$$

where for any $\lambda \in \mathbb{R}^M$ and $a > 0$, $B_\infty^M(\lambda, c)$ denotes the l_∞-ball in $\mathbb{R}^M$ of radius a centered at λ. Next, we easily have

$$\begin{aligned} \int_{\lambda_j^*-c\eta\epsilon}^{\lambda_j^*+c\eta\epsilon} p_{\lambda_j^*}(\lambda_j) \log\left(\frac{p_{\lambda_j^*}(\lambda_j)}{\pi_0(\lambda_j)}\right) d\lambda_j &= \left(1 - \frac{|\lambda_j^*|}{2c\eta\epsilon}\right)_+ \log\left(\frac{1+\epsilon}{(1-\gamma)(1+\epsilon)+\gamma\eta\epsilon}\right) \\ &\quad + \min\left(1, \frac{|\lambda_j^*|}{2c\eta\epsilon}\right) \log\left(\frac{1+\epsilon}{\gamma\eta\epsilon}\right). \end{aligned}$$

Since $0 < \gamma, \eta, \epsilon < 1$ we have that

$$\int_{\lambda_j^*-c\eta\epsilon}^{\lambda_j^*+c\eta\epsilon} p_{\lambda_j^*}(\lambda_j) \log \frac{p_{\lambda_j^*}(\lambda_j)}{\pi_0(\lambda_j)} d\lambda_j \leqslant \begin{cases} \log\left(\frac{1+\epsilon}{\gamma\eta\epsilon}\right) & \text{if } \lambda_j^* \neq 0, \\ \frac{\gamma}{1-\gamma}, & \text{if } \lambda_j^* = 0, \end{cases} \quad (7.41)$$

where we have used the inequality $\log(1+x) \leqslant x$ true for any $x > -1$.

We now treat the normalizing constant C_N. We have

$$\begin{aligned} C_N &= \int_{B_\infty^M(0,c\eta\epsilon)} \prod_{j=1}^{M} \pi_0(x_j) dx + \int_{\Lambda_\epsilon \cap B_\infty^M(0,c\eta\epsilon)^c} \prod_{j=1}^{M} \pi_0(x_j) dx \\ &\leqslant \left(\frac{1-\gamma}{2c\eta\epsilon} + \frac{\gamma}{2c(1+\epsilon)}\right)^M (2c\eta\epsilon)^M + \left(\frac{\gamma}{2c(1+\epsilon)}\right)^M C(M)(c(1+\epsilon))^M \\ &\leqslant \left(1 - \gamma + \frac{\gamma\eta\epsilon}{1+\epsilon}\right)^M + \left(\frac{\gamma}{2}\right)^M C(M) \\ &\leqslant 1, \quad (7.42) \end{aligned}$$

where $C(M) = \frac{\sqrt{M+1}}{M! 2^{\frac{M}{2}}}$ is the volume of the unit l_1 ball in $\mathbb{R}^M$. Combining (7.40),(7.41) and (7.42) yields

$$K(P^*, \Pi) \leqslant M(\lambda^*) \log\left(\frac{1+\epsilon}{\gamma\eta\epsilon}\right) + \frac{\gamma}{1-\gamma}(M - M(\lambda^*)).$$

With our choice of the parameters $\eta = \gamma = \frac{1}{M}$, we get

$$K(P^*, \Pi) \leqslant M(\lambda^*) \log\left(M^2 \frac{1+\epsilon}{\epsilon}\right) + 2\left(1 - \frac{M(\lambda^*)}{M}\right).$$

□

We have the following result.

Theorem 7.6. *For all $\epsilon > 0$ and any loss function Q such that the associated risk A satisfies Assumption (**A**), the aggregate $\tilde{\lambda}_n$ obtained by the implementation of the mirror averaging algorithm on the set Λ_ϵ with the prior Π, where Λ_ϵ and Π are defined in Lemma 7.5, satisfies the inequality*

$$E_{n-1}A(\tilde{\lambda}_n) \leqslant \min_{\lambda\in\Lambda}\left\{A(\lambda) + \beta\frac{M(\lambda)}{n}\left[2(\log M) + \log\left(\frac{1+\epsilon}{\epsilon}\right) + 2\left(\frac{1}{M(\lambda)} - \frac{1}{M}\right)\right]\right\} + 2L_1 c\epsilon + S_1,$$

where

$$S_1 \triangleq \beta E_n \log\left(\mathbb{E}_{\tilde{P}_n} \exp\left[-\frac{Q_1(Z_n, \omega, \tilde{\lambda}_n)}{\beta}\right]\right),$$

and ω is a random variable with values in Λ_ϵ and distributed with the law $\tilde{P}_n$.

Proof. The proof is exactly the same as that of Theorem 7.1 until the inequality (7.18). It suffices then to apply Lemma 7.5, (7.16), (7.17) and (7.18) to get the result. □

For the Gaussian regression model with random design considered in Section 7.5, we proved in Lemma 7.4 that $S_1 \leqslant 0$ for $\beta \geqslant \overline{\beta} = 2(\sigma^2 + (c+1)^2 L^2)$. Thus, applying Theorem 7.6 to this case with $\beta = \overline{\beta}$ and $\epsilon = 1/n$, we obtain the following corollary:

Corollary 7.1. *Let the assumptions of Theorem 7.6 be satisfied. Take $\epsilon = 1/n$ and $\beta = \overline{\beta} = 2(\sigma^2 + (c+1)^2 L^2)$. Then, for any distribution P^X, the aggregate $\hat{f}_n = f_{\tilde{\lambda}_n}$ satisfies the following oracle inequality:*

$$E_f^n ||\hat{f}_n - f||^2 \leqslant \min_{\lambda\in\Lambda}\left\{||f_\lambda - f||^2 + \overline{\beta}\frac{M(\lambda)}{n}\left[2(\log M) + \log(n+1) + 2\left(\frac{1}{M(\lambda)} - \frac{1}{M}\right)\right]\right\} + \frac{2L_1 c}{n}.$$

7.7.2 Taking Π as a distribution on the number of nonzero parameters of the model

The definition of the prior distribution Π used here is similar to the definition of the prior distribution used in Section 7.3 where we considered D-convex aggregation. Fix $c > 0$ and $\epsilon > 0$. Define $\Lambda = B_1^M(0, c)$ and $\Lambda_\epsilon = B_1^M(0, c(1+\epsilon))$. Every element $\lambda \in \Lambda_\epsilon$ can be associated to the following triplet $(M(\lambda), J(\lambda), \overline{\lambda}_{J(\lambda)})$ where $\overline{\lambda}_{J(\lambda)}$ denotes the vector of $\mathbb{R}^{M(\lambda)}$ obtained by keeping only the nonzero components of λ, with the convention that the zero vector is associated to the null triplet $(0, 0, 0)$. From now on, we redefine, w.l.o.g.,

$$\Lambda_\epsilon = \left\{(s, I, \tilde{\lambda}) \,:\, 1 \leqslant s \leqslant M,\, I \subseteq \{1, \dots, M\},\, |I| = s,\, \tilde{\lambda} \in B_1^s(0, c(1+\epsilon))\right\} \cup \{(0,0,0)\}.$$

We also consider the similar representation for Λ. We introduce first some probability distributions. Define the probability distribution Π_1 admitting the density π_1 w.r.t the counting measure on $\{0, \dots, M\}$ defined as follows:

$$\pi_1(j) = \frac{1 - e^{-1}}{1 - e^{-(M+1)}} e^{-j},\ \forall j \in \{0, \dots, M\}.$$

For any $s \in \{1, \dots, M\}$, denote by $\mathcal{I}_s$ the set of all subsets of $\mathcal{E}^M$ of cardinality s. We have

$$|\mathcal{I}_s| = \binom{M}{s} \leqslant \left(\frac{eM}{s}\right)^s.$$

Denote by $\Pi_{2,s}$ the uniform probability measure on $\mathcal{I}_s$. Its density $\pi_{2,s}$ w.r.t. the counting measure on $\mathcal{I}_s$ is given by

$$\pi_{2,s}(J) = \frac{1}{\binom{M}{s}},\ \forall J \subseteq \{1, \dots, M\} \,:\, |J| = s.$$

For any $s \in \{1, \dots, M\}$, denotes by $\Pi_{3,s}$ the uniform probability distribution on $B_1^s(0, c(1+\epsilon))$. Its admits the density $\pi_{3,s}$ w.r.t. the Lebesgue measure on $\mathbb{R}^s$ defined as follows:

$$\pi_{3,s}(\lambda) = \frac{1}{C(s)(c(1+\epsilon))^s} \mathbb{1}_{B_1^s(0, c(1+\epsilon))}(\lambda),\quad \forall \lambda \in \mathbb{R}^s,$$

where $C(s) = \frac{\sqrt{s+1}}{s! 2^{\frac{s}{2}}}$ is the volume of the unit l_1 ball in $\mathbb{R}^s$.

We consider the prior distribution Π on Λ_ϵ as the product of the above three probability measures. Its density w.r.t. the corresponding dominating measure is given by:

$$\pi(\lambda) = \pi_1(M(\lambda)) \pi_{2,M(\lambda)}(J(\lambda)) \pi_{3,M(\lambda)}(\overline{\lambda}_{J(\lambda)}). \tag{7.43}$$

Lemma 7.6. *Consider the sets $\Lambda = B_1^M(0,c)$ and $\Lambda_\epsilon = B_1^M(0,c(1+\epsilon))$ for a constant $c \geqslant 1$ and the prior distribution Π defined in (7.43). Let Assumption (**A**) be satisfied on Λ. Then, for any $\lambda^* \in \Lambda$ and any ϵ such that $0 < \epsilon \leqslant 1$, there exists a probability distribution P^* supported on Λ_ϵ such that*

$$\begin{aligned} K(P^*, \Pi) &\leqslant M(\lambda^*)\left(1 + \log\left(\frac{eM}{M(\lambda^*)}\right) + \log\left(\frac{1+\epsilon}{\epsilon}\right)\right) \\ |\overline{A}(P^*) - A(\lambda^*)| &\leqslant L_1 c M(\lambda^*)\epsilon. \end{aligned}$$

Proof. Set $s = M(\lambda^*)$ and $J^* = J(\lambda^*)$. Let P^* be the probability distribution admitting the following density

$$p_{\lambda^*}(\lambda) = \mathbb{1}_{M(\lambda)=s}\mathbb{1}_{J(\lambda)=J^*}\frac{1}{C(s)(c\epsilon)^s}\mathbb{1}_{B_1^s(\overline{\lambda^*}_{J^*}, c\epsilon)}(\overline{\lambda}_{J^*}), \quad \forall \lambda \in \mathbb{R}^M.$$

We have

$$\left|\int_\Lambda A(\lambda)p_{\lambda^*}(\lambda)d\lambda - A(\lambda^*)\right| \leqslant L_1 c M(\lambda^*)\epsilon.$$

Next, we have

$$\begin{aligned} K(P^*, \Pi) &= \log\left(\frac{1}{\pi_1(s)}\right) + \log\left(\frac{1}{\pi_{2,s}(J^*)}\right) + s\log\left(\frac{1+\epsilon}{\epsilon}\right) \\ &\leqslant s + s\log\left(\frac{eM}{s}\right) + s\log\left(\frac{1+\epsilon}{\epsilon}\right). \end{aligned}$$

This yields the result. □

We have the following result.

Theorem 7.7. *For all $\epsilon > 0$ and any loss function Q such that the associated risk A satisfies Assumption (**A**), the aggregate $\tilde{\lambda}_n$ obtained by the implementation of the mirror averaging algorithm on the set Λ_ϵ with the prior distribution Π, where Λ_ϵ and Π are defined in Lemma 7.6, satisfies the inequality*

$$E_{n-1}A(\tilde{\lambda}_n) \leqslant \min_{\lambda\in\Lambda}\left\{A(\lambda) + L_1 c M(\lambda)\epsilon + \beta\frac{M(\lambda)}{n}\left[1 + \log\left(\frac{eM}{M(\lambda)}\right) + \log\left(\frac{1+\epsilon}{\epsilon}\right)\right]\right\} + S_1,$$

where

$$S_1 \triangleq \beta E_n \log\left(\mathbb{E}_{\tilde{P}_n}\exp\left[-\frac{Q_1(Z_n, \omega, \tilde{\lambda}_n)}{\beta}\right]\right),$$

and ω is a random variable with values in Λ_ϵ and distributed with the law $\tilde{P}_n$.

Proof. The proof is exactly the same as that of Theorem 7.1 until the inequality (7.18). It suffices then to apply Lemma 7.6, (7.16), (7.17) and (7.18) to get the result. □

7.8 Appendix

Definition 7.2. *A function $T : \mathbb{R}^M \to \mathbb{R}$ is exponentially concave if the composite function $\exp \circ T$ is concave.*

Proposition 4.1 of [57] gives a simple sufficient condition for exponential concavity

Proposition 7.1. *Let g be a function twice differentiable on Λ^M with gradient $\nabla g(\lambda)$ and Hessian matrix $\nabla^2 g(\lambda)$, $\lambda \in \Lambda$. If there exists $\beta > 0$ such that for any $\lambda \in \Lambda$, the matrix*

$$\beta \nabla^2 g(\lambda) - \nabla g(\lambda)(\nabla g(\lambda))^T,$$

is semi-positive, then $-g(.)/\beta$ is exponentially concave on Λ.

Proof. Since g is a twice differentiable on Λ, then $\exp(-g(\cdot)/\beta)$ is also twice differentiable with Hessian matrix

$$\mathcal{H}(\lambda) = \frac{1}{\beta} \exp\left(-\frac{g(\lambda)}{\beta}\right) \left[\frac{\nabla g(\lambda)(\nabla g(\lambda))^T}{\beta} - \nabla^2 g(\lambda)\right].$$

For any $x \in \mathbb{R}^M$, $\lambda \in \Lambda$, we have

$$x^T \mathcal{H}(\lambda) x = \frac{1}{\beta} \exp\left(-\frac{g(\lambda)}{\beta}\right) \left[\frac{(x^T \nabla g(\lambda))^2}{\beta} - x^T \nabla^2 g(\lambda) x\right] \leqslant 0.$$

Hence $\exp\left(-\frac{g(\cdot)}{\beta}\right)$ has a negative semi-definite Hessian and is therefore concave. □

Let prove now Lemma 7.4.

Proof. Fix $\lambda' \in \Lambda$, consider the mapping

$$\lambda \to E \exp(-Q_1(Z, \lambda, \lambda')/\beta) = E \exp\left(-\frac{1}{\beta}\left[(Y - f_\lambda(X))^2 - (Y - f_{\lambda'}(X))^2\right]\right).$$

We have

$$\begin{aligned} E \exp(-Q_1(Z, \lambda, \lambda')/\beta) &= E \exp\left(-\frac{1}{\beta}\left[(Y - f_\lambda(X))^2 - (Y - f_{\lambda'}(X))^2\right]\right) \\ &= E \exp\left(-\frac{1}{\beta}\left[-2\xi(U(X,\lambda) - U(X,\lambda')) + U^2(X,\lambda) - U^2(X,\lambda')\right]\right), \end{aligned}$$

where $U(X, \lambda) \triangleq f(X) - f_\lambda(X)$. Since $|2(U(X,\lambda) - U(X,\lambda'))| = 2|(f_{\lambda - \lambda'}(X)| \leqslant 4Lc$ and ξ is gaussian $\mathcal{N}(0, \sigma^2)$, taking the expectation conditionally on X first, we get

$$E \exp(-Q_1(Z, \lambda, \lambda')/\beta) = \psi_\beta(\lambda, \lambda'),$$

where

$$\psi_\beta(\lambda,\lambda') \triangleq E\exp\left(\frac{2\sigma^2}{\beta^2}(f_{\lambda-\lambda'}(X))^2 - \frac{1}{\beta}[U^2(X,\lambda) - U^2(X,\lambda')]\right). \tag{7.44}$$

Clearly

$$\psi_\beta(\lambda,\lambda) = 1,\ \forall\lambda\in\Lambda. \tag{7.45}$$

For any $x\in\mathbb{R}^d$ and $\lambda'\in\Lambda$ fixed, consider the function

$$\lambda \to \tilde{Q}(x,\lambda,\lambda') \triangleq \left(-1+\frac{2\sigma^2}{\beta}\right)(f_\lambda(x))^2 - \frac{4\sigma^2}{\beta}(f_\lambda(x))(f_{\lambda'}(x)) + 2f(x)f_\lambda(x).$$

We give now a sufficient condition for $\tilde{Q}(x,\cdot,\lambda')$ to be exponentially concave. Denote $\gamma = 1-\frac{2\sigma^2}{\beta}$ and $H(x) = (f_1,\cdots,f_M)(x)^T$, we have

$$\nabla_\lambda\tilde{Q}(x,\lambda,\lambda') = \left(-2\gamma(H^T(x)\lambda) - \frac{4\sigma^2}{\beta}(H^T(x)\lambda') + 2f(x)\right)H(x),$$

$$\nabla^2_{\lambda\lambda}\tilde{Q}(x,\lambda,\lambda') = -2\gamma H(x)H^T(x).$$

By assumption $||f||_\infty \leqslant L$ and since $\lambda\in\Lambda$, we have $|H^T(x)\lambda| \leqslant Lc$, $|H^T(x)\lambda'| \leqslant Lc$. If

$$\beta \geqslant \beta_0 \triangleq 2(\sigma^2 + ((c+1)L)^2), \tag{7.46}$$

then $\beta\nabla^2_{\lambda\lambda}\tilde{Q}(x,\lambda,\lambda') - (\nabla_\lambda\tilde{Q}(x,\lambda,\lambda'))(\nabla_\lambda\tilde{Q}(x,\lambda,\lambda'))^T$ is semi-definite. Proposition 7.1 ensures that $\tilde{Q}(x,\cdot,\lambda')/\beta$ is exponentially concave in λ. So is $\psi_\beta(\cdot,\lambda')/\beta$ for any $\lambda'\in\Lambda$ fixed. Consider a random variable ω with values in Λ and distributed as $\tilde{P}_n$. Denote $\mathbb{E}_{\tilde{P}_n}$ the distribution w.r.t. $\tilde{P}_n$, we have $\tilde{\lambda}_n = \mathbb{E}_{\tilde{P}_n}[\omega]$. We have for $\beta\geqslant\beta_0$

$$\begin{aligned}
S_n &= E_n\log\left(\mathbb{E}_{\tilde{P}_n}\exp\left[-\frac{Q_1(Z_n,\omega,\tilde{\lambda}_n)}{\beta}\right]\right)\\
S_n &\leqslant E_{n-1}\log\left(E\mathbb{E}_{\tilde{P}_n}\exp\left[-\frac{Q_1(Z_n,\omega,\tilde{\lambda}_n)}{\beta}\right]\right)\\
&= E_{n-1}\log\left(\mathbb{E}_{\tilde{P}_n}E\exp\left[-\frac{Q_1(Z_n,\omega,\tilde{\lambda}_n)}{\beta}\right]\right)\\
&= E_{n-1}\log\left(\mathbb{E}_{\tilde{P}_n}\psi_\beta(\omega,\tilde{\lambda}_n)\right)\\
&\leqslant E_{n-1}\log\left(\psi_\beta(\mathbb{E}_{\tilde{P}_n}[\omega],\tilde{\lambda}_n)\right) = E_{n-1}\log\left(\psi_\beta(\tilde{\lambda}_n,\tilde{\lambda}_n)\right) = 0.
\end{aligned}$$

In the second line, we used Jensen inequality w.r.t. the expectation E of the random variable Z_n. In the third line, we used Fubini's Theorem. In the fourth line, we used the fact that Z_n and $\tilde{\lambda}_n$ are independent and the definition of $\psi_\beta(\lambda,\lambda')$ (Recall that $\tilde{\lambda}_n$ is a measurable

function of $Z_1, \ldots, Z_{n-1}$). In the last line, we used the fact that for any fixed $\lambda' \in \Lambda$, $\psi_\beta(\cdot, \lambda')$ is concave if $\beta \geqslant \beta_0$ and (7.45).

□

Chapter 8

Oracle inequalities for the L^π norm in a density estimation problem

We study the Goldenshluger and Goldenshluger-Lepski model selection procedures in a density estimation framework. We derive oracle inequalities for the L^π risk with $1 \leqslant \pi \leqslant \infty$. Then, we exploit these oracle inequalities to propose minimax rate adaptive procedures.

8.1 Introduction

Consider i.i.d. random vectors $X_1, \dots, X_n$ with values in a Borel subset $\mathcal{X}$ of $\mathbb{R}^d$ having an unknown common probability density $f \in L^\pi(\mathcal{X})$ that we want to estimate, where $\pi \geqslant 1$. For a function $g \in L^\pi(\mathcal{X})$ and $\pi < \infty$, define $||g||_\pi = \left(\int_{\mathcal{X}} |g(x)|^\pi dx\right)^{1/\pi}$. If $\pi = \infty$, set $||g||_\infty = \operatorname{ess\,sup}_{x\in\mathcal{X}} |g(x)|$. For an estimator $\hat{f}$ of f based on the sample $\mathbb{X}_n = (X_1, \dots, X_n)$, define the L^π risk $E_f^{\otimes n}(||\hat{f} - f||_\pi^\alpha)$, where $\alpha > 0$ and $E_f^{\otimes n}$ denotes the expectation w.r.t. the distribution $P_f^{\otimes n}$ of $\mathbb{X}_n$. Consider the dictionary $\mathcal{F} = \{f_1, \dots, f_M\}$ where $M \geqslant 2$ and f_j, $1 \leqslant j \leqslant M$, are some estimators of the density f. We study the model selection type aggregation problem: given $\mathbb{X}_n$ we want to construct a new estimator $\tilde{f}_n$ of f, called aggregate, which is approximately at least as good as the best estimator among $f_1, \dots, f_M$, in the sense that it satisfies the oracle inequality

$$E_f^{\otimes n}(||\tilde{f}_n - f||_\pi^\alpha) \leqslant C \min_{1\leqslant j\leqslant M} E_f^{\otimes n}(||f_j - f||_\pi^\alpha) + \Delta(n, M, \pi, \alpha), \tag{8.1}$$

where $C \geqslant 1$ is an absolute constant and the remainder term $\Delta(n, M, \pi, \alpha)$ called the rate of aggregation does not depend on f and should be as small as possible. This problem is called (MS) aggregation problem.

The literature on aggregation in regression and Gaussian white noise models is extensive. We mention here only some recent works [15, 57, 68, 70, 97, 106] and the references cited therein. Most of the results concern the L^2 risk. Hengartner and Wegkamp [52] derive an oracle inequality of the form (8.1) in a regression model under the L^π risk with $1 \leqslant \pi \leqslant 2$. In the Gaussian white noise model, Goldenshluger [46] establishes an oracle inequality for the L^π risk with $1 \leqslant \pi \leqslant \infty$. The aggregation procedure proposed in [52] and [46] is based on the minimization of the supremum of a linear functional on an appropriate set of functions .

Aggregation of density estimators has been considered in [8, 18, 57, 92, 109, 115] with the Kullback-Leibler divergence and L^2 risks. Devroye and Lugosi [28] and Birgé [7] obtained results under the L^1 risk. Optimality of the rates of aggregation in the sense of [97] have been proved in [89] for the L^2 risk and in [65] for the Kullback-Leibler divergence and the L^1 risk.

In this chapter, we extend the procedure of Hengartner and Wegkamp [52], Goldenshluger [46] and Goldenshluger and Lepski [47] to the problem of aggregation of density estimators. For $\pi = 1$, the procedure of Devroye and Lugosi [28] appears as a particular case of our procedure. We will discuss this point in more detail in Section 8.2. The procedure of Goldenshluger and Lepski is related to Lepski's method. This procedure is adapted to a specific class of linear estimators including the linear wavelet estimators. Goldenshluger and Lepski

[47] considered the white noise model and established a model selection oracle inequality for the sup-norm. In Section 8.3, we consider the density estimation framework and we derive a model-selection oracle inequality for the sup-norm. Note that the adaptation of the result of Goldenshluger and Lepski to the density estimation problem is straightforward. We decided to write it down since we exploit this result later. In Section 8.4, we exploit the results of Sections 8.2 and 8.3 to build minimax rate adaptive density estimators for the L^π norm, with $1 \leqslant \pi \leqslant 2$ and the sup-norm.

8.2 The Goldenshluger procedure

In this section, $f_1, \ldots, f_M$ are deterministic functions.

8.2.1 The aggregation procedure

The procedure is based on estimation of linear functionals over an appropriately chosen set of functions Ω on $\mathcal{X}$ with values in $\mathbb{R}$. Consider the linear functional

$$l_f(\omega) = \int_{\mathcal{X}} \omega(t) f(t) dt, \; \omega \in \Omega.$$

A natural estimator of $l_f(\omega)$ is

$$\hat{l}_f(\omega) = \frac{1}{n} \sum_{i=1}^{n} \omega(X_i).$$

Consider the following functionals

$$l_{f_j}(\omega) = \int_{\mathcal{X}} \omega(t) f_j(t) dt, \quad j = 1, \ldots, M,$$

and the differences between these two quantities

$$\Delta_j(\omega) = \hat{l}_f(\omega) - l_{f_j}(\omega), \quad j = 1, \ldots, M.$$

The aggregation procedure is defined as follows. Set

$$\hat{M}_j = \sup_{\omega \in \Omega} \frac{|\Delta_j(\omega)|}{||\omega||_q}, \quad j = 1, \ldots, M, \tag{8.2}$$

where $\frac{1}{\pi} + \frac{1}{q} = 1$, with the convention $q = 1$ if $\pi = \infty$. Consider

$$\hat{j} = \arg \min_{1 \leqslant j \leqslant M} \hat{M}_j. \tag{8.3}$$

Define the aggregate $\tilde{f}_n$ by

$$\tilde{f}_n = f_{\hat{j}}. \tag{8.4}$$

Since $E_f^{\otimes n}(\Delta_j(\omega)) = l_{f-f_j}$, $\hat{M}_j$ is the empirical estimate of the supremum of $|l_{f-f_j}(\omega)|/||\omega||_q$ over Ω. If f_j is close to f in the L^π norm, then by Hölder's inequality, $\hat{M}_j$ should be small.

The choice of the set of functions Ω is crucial for the performance of the procedure. Following Goldenshluger [46], we consider the following choice. Given the dictionary $\mathcal{F} = \{f_1, \ldots, f_M\}$, define the set $\mathcal{G} = \{g : \mathcal{X} \to \mathbb{R} | g = g_{i,j} = f_i - f_j, \forall i \neq j\}$. Let $\pi \in [1, \infty)$. From now on, we consider the following set of functions

$$\Omega = \{\omega | \omega(\cdot) = \omega_g(\cdot) = |g(\cdot)|^{\pi-1} ||g||_\pi^{-\pi+1} \text{sign}\{g(\cdot)\},\ g \in \mathcal{G}\}. \tag{8.5}$$

It is easy to see that for all $g \in \mathcal{G}$, there exists $\omega_g \in \Psi$ such that

$$\int_{\mathcal{X}} \frac{\omega_g(t)}{||\omega_g||_q} g(t) dt = ||g||_\pi. \tag{8.6}$$

Note that $||\omega||_q = 1$ for all ω in the set Ω. So our aggregation procedure is easily implementable since it amounts to the estimation of empirical linear functionals and a minimization-maximization step over a finite set.

The aggregation procedure considered by Devroye and Lugosi [28] is the minimum distance estimate initially proposed by Yatracos [111]:

$$\hat{j}^Y = \arg \min_{1 \leqslant j \leqslant M} \sup_{A \in \mathcal{A}} \left| \int_A f_j - \mu_n(A) \right|, \quad \tilde{f}_n^Y = f_{\hat{j}^Y},$$

where $\mu_n(A) = (1/n) \sum_{i=1}^n \mathbb{1}_{X_i \in A}$ is the empirical measure of an event A based on the data $X_1, \ldots, X_n$, $\mathbb{1}_B$ is the indicator function of the event B and $\mathcal{A}$ is the Yatracos class

$$\mathcal{A} = \left\{ A_{i,j} \stackrel{\triangle}{=} \{x : f_i(x) > f_j(x)\} ; 1 \leqslant i, j \leqslant M \right\}.$$

With this estimation procedure, Devroye and Lugosi [28] proved an oracle inequality of the form (8.1) under the L^1 risk. Consider now our procedure under the L^1 risk with the set of functions of (8.5). In this case, any function $\omega \in \Omega$ is of the form $\omega(x) = 2\mathbb{1}_{f_i(x) > f_j(x)} - 1$ for $i, j \in \{1, \ldots, M\}$. Thus our estimation procedure can be rewritten as

$$\hat{j} = \arg \min_{1 \leqslant j \leqslant M} \sup_{A \in \mathcal{A}} \left| 2(\int_A f_j - \mu_n(A)) + 1 - \int_{\mathcal{X}} f_j \right|, \quad \tilde{f}_n = f_{\hat{j}},$$

If the functions f_j, $j = 1, \ldots, M$, are densities, then our procedure under the L^1 risk is exactly the Yatracos procedure. So we can say that our procedure is a generalization of the Yatracos procedure to the L^π risk with $1 \leqslant \pi < \infty$.

8.2.2 Oracle inequalities for the L^π risk with $1 \leqslant \pi < \infty$

In this section, we prove the main results of the chapter. These are oracle inequalities in expectation and in probability for the density aggregation problem. The next theorem states the main oracle inequality in expectation under the L^π risk for the aggregate $\tilde{f}_n$.

Theorem 8.1. *Fix $\pi \in [1, \infty)$ and $\alpha > 0$. Consider a dictionary $\mathcal{F} = \{f_1, \ldots, f_M\}$. Let Ω be the finite set of functions (8.5) associated with $\mathcal{F}$. Assume that $b_\Omega \stackrel{\triangle}{=} \max_{\omega \in \Omega} ||\omega||_\infty < \infty$. Then the aggregate $\tilde{f}_n$ defined in (8.4) satisfies the oracle inequality*

$$E_f^{\otimes n}\left(||\tilde{f}_n - f||_\pi^\alpha\right) \leqslant C_1 \min_{1 \leqslant j \leqslant M} ||f_j - f||_\pi^\alpha + C_2\left(b_\Omega \frac{\log(1+|\Omega|)}{n}\right)^\alpha + C_3\left(\sigma_\Omega^2 \frac{\log(1+|\Omega|)}{n}\right)^{\frac{\alpha}{2}}, \tag{8.7}$$

where $\sigma_\Omega^2 \stackrel{\triangle}{=} \max_{\omega \in \Omega} \int_{\mathcal{X}} \omega^2(x) f(x) dx < \infty$, $|\Omega|$ denotes the cardinality of Ω, $C_1 = 3^\alpha(2^{\alpha \vee 1 - 1})$, $C_2 = 8^\alpha(2^{\alpha \vee 1 - 1})$ and $C_3 = 24^\alpha(2^{\alpha \vee 1 - 1})$.

Proof. For all $1 \leqslant j \leqslant M$ we have

$$\begin{aligned} \hat{M}_j &= \max_{\omega \in \Omega} |\Delta_j(\omega)| \\ &\leqslant \max_{\omega \in \Omega} (|\hat{l}_f(\omega) - l_f(\omega)| + |l_f(\omega) - l_{f_j}(\omega)|) \\ &\leqslant ||f_j - f||_\pi + \max_{\omega \in \Omega} |\hat{l}_f(\omega) - l_f(\omega)|, \end{aligned} \tag{8.8}$$

where we have used the Hölder inequality in the third line. Define $\jmath^* = \arg\min_{1 \leqslant j \leqslant M} ||f_j - f||_\pi$. We have

$$\begin{aligned} ||\tilde{f}_n - f||_\pi &= ||f_{\hat{\jmath}} - f||_\pi \\ &\leqslant ||f_{\hat{\jmath}} - f_{\jmath^*}||_\pi + ||f_{\jmath^*} - f||_\pi \\ &\leqslant |l_{f_{\hat{\jmath}} - f_{\jmath^*}}(\psi_{f_{\hat{\jmath}} - f_{\jmath^*}})| + ||f_{\jmath^*} - f||_\pi \\ &\leqslant \hat{M}_{\hat{\jmath}} + \hat{M}_{\jmath^*} + ||f_{\jmath^*} - f||_\pi \\ &\leqslant 2\hat{M}_{\jmath^*} + ||f_{\jmath^*} - f||_\pi \\ &\leqslant 3||f_{\jmath^*} - f||_\pi + 2 \max_{\omega \in \Omega} |\hat{l}_f(\omega) - l_f(\omega)|, \end{aligned}$$

where we have used (8.6) in the second line and (8.8) in the last line. Consider first the case $\alpha \geqslant 1$. We have by convexity of the application $x \to x^\alpha$ that

$$||\tilde{f}_n - f||_\pi^\alpha \leqslant C_1 ||f_{\jmath^*} - f||_\pi^\alpha + 2^{2\alpha - 1} \max_{\omega \in \Omega} \left(|\hat{l}_f(\omega) - l_f(\omega)|^\alpha\right). \tag{8.9}$$

Taking the expectation in the above inequality, we get that

$$E_f^{\otimes n}\left(||\tilde{f}_n - f||_\pi^\alpha\right) \leqslant C_1||f_{\jmath^*} - f||_\pi^\alpha + 2^{2\alpha-1}E_f^{\otimes n}\left(\max_{\omega\in\Omega}\left(|\hat{l}_f(\omega) - l_f(\omega)|^\alpha\right)\right).$$

We rewrite this result as follows:

$$\begin{aligned} E_f^{\otimes n}\left(||\tilde{f}_n - f||_\pi^\alpha\right) \leqslant & C_1 \min_{1\leqslant j\leqslant M}||f_\jmath - f||_\pi^\alpha + \\ & + 2^{2\alpha-1}E_f^{\otimes n}\left(\max_{\omega\in\Omega}\left(\left|\frac{1}{n}\sum_{i=1}^n \omega(X_i) - \int_{\mathcal{X}}\omega(t)f(t)dt\right|\right)^\alpha\right). \end{aligned} \tag{8.10}$$

Now we study the expectation in the right hand term. Define the random variables

$$Y_{i,\omega} = \omega(X_i) - \int_{\mathcal{X}}\omega(t)f(t)dt,$$

and

$$Z_\omega = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_{i,\omega}, \tag{8.11}$$

for any $1 \leqslant i \leqslant n$ and $\psi \in \Omega$. The $Y_{i,\omega}$'s are independent, zero mean random variables of finite variance $\sigma_\omega^2 = \int_{\mathcal{X}}\omega(t)^2 f(t)dt - \left(\int_{\mathcal{X}}\omega(t)f(t)dt\right)^2$ and bounded by $2||\omega||_\infty$. For any $\omega \in \Omega$, we can apply the Bernstein inequality and we get that

$$P_f^{\otimes n}\left(|Z_\omega| > t\right) \leqslant 2\exp\left(-\frac{t^2}{2\sigma_\omega^2 + 2t||\omega||_\infty/(3\sqrt{n})}\right),$$

for any $t > 0$. Define the quantity $b_\omega = ||\omega||_\infty/(3\sqrt{n})$. We have that

$$P_f^{\otimes n}\left(|Z_\omega| > t\right) \leqslant \begin{cases} 2\exp\left(-\frac{t^2}{4\sigma_\omega^2}\right), & \text{if } t < \sigma_\omega^2/b_\omega, \\ 2\exp\left(-\frac{t}{4b_\omega}\right), & \text{if } t \geqslant \sigma_\omega^2/b_\omega. \end{cases}$$

Define the random variables $T_\omega = Z_\omega \mathbb{I}_{|Z_\omega| > \sigma_\omega^2/b_\omega}$ and $T'_\omega = Z_\omega \mathbb{I}_{|Z_\omega| \leqslant \sigma_\omega^2/b_\omega}$. For all $t > 0$ we have

$$P_f^{\otimes n}\left(|T_\omega| > t\right) \leqslant 2\exp\left(-\frac{t}{4b_\omega}\right), \quad P_f^{\otimes n}\left(|T'_\omega| > t\right) \leqslant 2\exp\left(-\frac{t^2}{4\sigma_\omega^2}\right).$$

Define the function $h_\nu(x) = \exp(x^\nu) - 1$, where $\nu > 0$. This function is clearly convex for any $\nu > 0$. We have that

$$E_f^{\otimes n}\left(h_{1/\alpha}\left(\left(\frac{|T_\omega|}{12b_\omega}\right)^\alpha\right)\right) = \int_0^\infty e^t P_f^{\otimes n}(|T_\omega| > 12b_\omega t)dt \leqslant 1,$$

where we have used Fubini's Theorem in the first equality. Since the function $h_{1/\alpha}$ is convex and nonnegative, we have that

$$\begin{aligned} h_{1/\alpha}\left(E_f^{\otimes n}\left(\max_{\psi\in\Omega}\left(\frac{|T_\omega|}{12b_\omega}\right)^\alpha\right)\right) &\leqslant E_f^{\otimes n}\left(h_{1/\alpha}\left(\max_{\omega\in\Omega}\left(\frac{|T_\omega|}{12b_\omega}\right)^\alpha\right)\right) \\ &\leqslant E_f^{\otimes n}\left(\sum_{\omega\in\Omega} h_{1/\alpha}\left(\left(\frac{|T_\omega|}{12b_\omega}\right)^\alpha\right)\right) \\ &\leqslant |\Omega|, \end{aligned}$$

where we have used the Jensen inequality in the first line. Since the function $h_{1/\alpha}^{-1}(x) = (\log(1+x))^\alpha$ is increasing, we have that

$$\begin{aligned} E_f^{\otimes n}\left(\max_{\omega\in\Omega}\left(\frac{|T_\omega|}{12b_\omega}\right)^\alpha\right) &\leqslant (\log(1+|\Omega|))^\alpha \\ E_f^{\otimes n}\left(\max_{\omega\in\Omega}\left(|T_\omega|\right)^\alpha\right) &\leqslant (12b_\Omega \log(1+|\Omega|))^\alpha . \end{aligned} \tag{8.12}$$

By the same arguments with the function $h_{2/\alpha}$, we prove that

$$E_f^{\otimes n}\left(\max_{\omega\in\Omega}\left(|T_\omega|\right)^\alpha\right) \leqslant \left(12\sigma_\Omega^2 \log(1+|\Omega|)\right)^{\frac{\alpha}{2}} . \tag{8.13}$$

Combining (8.10), (8.12) and (8.13), we get the result for $\alpha \geqslant 1$.

Consider now the case $0 < \alpha < 1$. The oracle inequality is an immediate consequence of the case $\alpha = 1$ by the Jensen inequality and the inequality $(x+y)^\alpha \leqslant x^\alpha + y^\alpha$ for all $x, y \geqslant 0$. □

Now we state an oracle inequality in probability. Theorem 8.2 below yields an analog of Theorem 8.1 with an oracle inequality "in probability".

Theorem 8.2. *Fix $\pi \in [1,\infty)$ and $\alpha > 0$. Consider a dictionary $\mathcal{F} = \{f_1, \ldots, f_M\}$. Let Ω be the finite set of functions (8.5) associated with $\mathcal{F}$. Assume that $b_\Omega \stackrel{\triangle}{=} \max_{\omega\in\Omega} ||\omega||_\infty < \infty$. Fix $\delta \in (0,1)$. Then on an event A_δ of probability $P_f^{\otimes n}(A_\delta) \geqslant 1-\delta$, the aggregate $\tilde{f}_n$ defined in (8.4) satisfies the following oracle inequality*

$$\begin{aligned} ||\tilde{f}_n - f||_\pi^\alpha \leqslant & C_1 \min_{1\leqslant j\leqslant M} ||f_j - f||_\pi^\alpha \\ & + C_2\left(2b_\Omega \frac{\log\left(2|\Omega|/\delta\right)}{3n} + \sigma_\Omega\sqrt{\frac{\log\left(2|\Omega|/\delta\right)}{n}}\right)^\alpha , \end{aligned}$$

where $\sigma_\Omega^2 \stackrel{\triangle}{=} \max_{\omega\in\Omega}\int_{\mathcal{X}} \omega^2(x) f(x) dx < \infty$, $C_1 = 3^\alpha(2^{\alpha\vee 1 - 1})$ and $C_2 = 2^\alpha(2^{\alpha\vee 1-1})$.

Proof. The proof is similar to that of Theorem 8.1 until the inequality (8.9). Define the event

$$A_\delta = \left\{ \max_{\omega \in \Omega} |Z_\omega| \leqslant t_\delta \right\},$$

where the random variables are defined in (8.11) and the quantity t_δ is specified later.

$$\begin{aligned} P_f^{\otimes n}(A_\delta^c) &\leqslant \sum_{\omega \in \Omega} P_f^{\otimes n}(|Z_\omega| > t_\delta) \\ &\leqslant \sum_{\omega \in \Omega} P_f^{\otimes n}(|Z_\omega| > t_\delta, |Z_\omega| > \sigma_\omega^2/b_\omega) + P_f^{\otimes n}(|Z_\omega| > t_\delta, |Z_\omega| \leqslant \sigma_\omega^2/b_\omega), \end{aligned}$$

where $\sigma_\omega^2 = \int_{\mathcal{X}} \omega(t)^2 f(t) dt - \left(\int_{\mathcal{X}} \omega(t) f(t) dt \right)^2$ and $b_\omega = ||\omega||_\infty / (3\sqrt{n})$. Thus by the Bernstein inequality, we have

$$P_f^{\otimes n}(A_\delta^c) \leqslant 2 \sum_{\omega \in \Omega} \exp\left(-\frac{t_\delta}{4 b_\omega} \right) + \exp\left(-\frac{t_\delta^2}{4\sigma_\omega^2} \right).$$

Take $t_\delta = \frac{4 b_\Omega}{3\sqrt{n}} \log\left(\frac{2|\Omega|}{\delta}\right) + 2\sigma_\Omega \sqrt{\log\left(\frac{2|\Omega|}{\delta}\right)}$ to get the result. □

The following two corollaries are immediate consequences of Theorems 8.1 and 8.2 respectively if we assume that $||f||_\infty < \infty$.

Corollary 8.1. *Let the assumptions of Theorem 8.1 hold. Assume that $||f||_\infty < \infty$. Then the aggregate $\tilde{f}_n$ defined in (8.4) satisfies the oracle inequality*

$$\begin{aligned} E_f^{\otimes n}\left(||\tilde{f}_n - f||_\pi^\alpha \right) \leqslant & C_1 \min_{1 \leqslant j \leqslant M} ||f_j - f||_\pi^\alpha \\ & + C_2 \left(Q_2(\pi) \frac{\log M}{n} \right)^\alpha + C_3 \left(Q_1(\pi) \sqrt{||f||_\infty \frac{\log M}{n}} \right)^\alpha, \end{aligned} \tag{8.14}$$

where

$$Q_1(\pi) = \max_{i \neq j} \frac{||f_j - f_i||_{2\pi-2}^{\pi-1}}{||f_j - f_i||_\pi^{\pi-1}}, \tag{8.15}$$

$$Q_2(\pi) = \max_{i \neq j} \frac{||f_j - f_i||_\infty^{\pi-1}}{||f_j - f_i||_\pi^{\pi-1}}, \tag{8.16}$$

$C_1 = 3^\alpha (2^{\alpha \vee 1 - 1})$, $C_2 = 16^\alpha (2^{\alpha \vee 1 - 1})$ *and* $C_3 = (24\sqrt{2})^\alpha (2^{\alpha \vee 1 - 1})$.

Proof. For the set of functions given in (8.5), we have $|\Omega| = M(M-1)/2$. Clearly we have that $b_\Omega = Q_2(\pi)$ and $\sigma_\Omega^2 \leqslant ||f||_\infty Q_1(\pi)$. Theorem 8.1 yields the result. □

In a similar way, we get the following corollary of Theorem 8.2.

Corollary 8.2. *Let the assumptions of Corollary 8.1 hold. Fix $\delta \in (0,1)$. Then on an event A_δ of probability $P_f^{\otimes n}(A_\delta) \geqslant 1-\delta$, the aggregate $\tilde{f}_n$ defined in (8.4) satisfies the following oracle inequality*

$$||\tilde{f}_n - f||_\pi^\alpha \leqslant C_1 \min_{1\leqslant j\leqslant M} ||f_j - f||_\pi^\alpha + C_2 \left(2Q_2(\pi)\frac{\log(M^2/\delta)}{3n} + Q_1(\pi)\left(||f||_\infty \frac{\log(M^2/\delta)}{n}\right)^{1/2}\right)^\alpha, \tag{8.17}$$

where $Q_j(\pi), j=1,2$ are defined in Corollary 8.1 and $C_j, j=1,2$ in Theorem 8.2.

Proof. For the set of functions given in (8.5), we have $|\Omega| = M(M-1)/2$. Clearly we have that $b_\Psi = Q_2(\pi)$ and $\sigma_\Omega^2 \leqslant ||f||_\infty Q_1(\pi)$. Theorem 8.2 yields the result. □

We observe that if $\pi = 1$ our results coincide with those obtained by Devroye and Lugosi [28] for density aggregation under the L^1 risk. Our results are more general since they cover the L^π risk for $1 \leqslant \pi < \infty$. It is interesting to compare our results with Goldenshluger [46] where aggregation in the Gaussian white noise model has been considered. As in [46], the aggregation rates in Corollaries 8.1 and 8.2 contain the term $Q_1(\pi)\sqrt{(\log M)/n}$. For $1 \leqslant \pi \leqslant 2$, we have $Q_1(\pi) \leqslant 1$. For $\pi > 2$, this term depends explicitly on the dictionary and can be very large. In the density aggregation problem, the rate of aggregation contain an additionally term $Q_2(\pi)(\log M)/n$ which does not appear in the Gaussian white noise aggregation problem [46]. We can see that the quantity $Q_2(\pi)$ depends explicitly on the dictionary and can be large even for $1 \leqslant \pi \leqslant 2$. In the next section, we prove that these bounds cannot be improved.

8.2.3 Lower bounds

We show that the oracle inequalities (8.14) and (8.17) cannot be improved in a minimax sense. We recall that the Kullback-Leibler divergence between two probability measures P and Q is defined by

$$K(P,Q) = \begin{cases} \int \log(\frac{dP}{dQ})dP, & \text{if } P \ll Q \\ +\infty, & \text{elsewhere.} \end{cases}$$

The following lemma can be proved by combining Theorems 2.2 and 2.5 in [98].

Lemma 8.1. *Let $w : \mathbb{R}^+ \to \mathbb{R}^+$ be a nondecreasing function such that $w(0) = 0$. Let $(r_n)_{n\in\mathbb{N}}$ be a sequence of positive numbers. Let $\mathcal{C}$ be a finite set of densities on $\mathcal{X}$ such that*

$|\mathcal{C}| \geqslant 2$,

$$||f - g||_\pi \geqslant cr_n > 0, \, \forall f, g \in \mathcal{C}, \, f \neq g,$$

for some $c > 0$. *If the Kullback-Leibler divergence* $K(P_f^{\otimes n}, P_g^{\otimes n})$ *between probability distributions* $P_f^{\otimes n}$ *and* $P_g^{\otimes n}$ *associated with the densities* f *and* g *satisfies*

$$\forall f, g \in \mathcal{C}, \, K(P_f^{\otimes n}, P_g^{\otimes n}) \leqslant \frac{1}{16} \log |\mathcal{C}|,$$

then,

$$\inf_{T_n} \sup_{f \in \mathcal{C}} E_f^{\otimes n} \left(w \left(r_n^{-1} ||T_n - f||_\pi \right) \right) \geqslant c_1,$$

where $\inf_{T_n}$ *denotes the infimum over all the estimators and* $c_1 > 0$ *is a constant.*

The following theorem implies that the upper bounds derived in Section 8.2 are the optimal rates of aggregation.

Theorem 8.3. *Let the integers* $M \geqslant 2$ *and* $n \geqslant 1$ *be such that* $M \log M \leqslant c_0 n$ *where* $c_0 \geqslant 4(1 + 2\log(4/3))$. *Fix* $\pi \in [1, \infty]$. *Let* $\mathcal{P}$ *be the set of all the probability densities* $f \in L^\pi(\mathbb{R}^d)$ *such that* $||f||_\infty \leqslant L$ *for some* $L > 0$. *Then there exist probability densities* $f_j \in L^\pi(\mathbb{R}^d)$, $j = 1, \ldots, M$, *such that for all estimators* T_n *of* f, *we have*

$$\sup_{f \in \mathcal{P}} E_f^{\otimes n} \left(w(r_n^{-1}(\|T_n - f\|_\pi^\alpha - \min_{1 \leqslant j \leqslant M} \|f_j - f\|_\pi^\alpha)) \right) \geqslant c,$$

for any integer $n \geqslant 1$, *for some constant* $c > 0$, *where*

$$r_n = \left(Q_1(\pi) \sqrt{L \frac{\log M}{n}} \vee Q_2(\pi) \frac{\log M}{n} \right)^\alpha,$$

and the quantities $Q_1(\pi)$ *and* $Q_2(\pi)$ *are defined in Corollary 1.*

Proof. Define the functions $\tilde{f}_0$ and $\tilde{g}$ on $\mathbb{R}$ by

$$\tilde{f}_0(x_1) = \frac{2L}{3} \mathbb{1}_{[0, \frac{3}{2L}]}(x_1), \quad \tilde{g}(x_1) = \frac{L}{3} \mathbb{1}_{[0, \frac{3}{4L}]}(x_1) - \frac{L}{3} \mathbb{1}_{[\frac{3}{4L}, \frac{3}{2L}]}(x_1),$$

where $x_1 \in \mathbb{R}$ and $\mathbb{1}_A$ denotes the indicator function of a set A. Consider the functions $\tilde{g}_j(x_1) = \tilde{g}(Mx_1 - 3(j-1)/2L)$, $1 \leqslant j \leqslant M$. Clearly for any $j \in \{1, \ldots, M\}$, the function $\tilde{g}_j$ is compactly supported on $[3(j-1)/(2LM), 3j/(2LM)]$. Define the functions f_0 and g_j on $\mathbb{R}^d$ by

$$f_0(x) = \tilde{f}_0(x_1) \prod_{k=2}^d \mathbb{1}_{[0,1]}(x_k), \quad g_j(x) = \tilde{g}_j(x_1) \prod_{k=2}^d \mathbb{1}_{[0,1]}(x_k),$$

where $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Consider now the set of functions

$$\mathcal{F} = \left(f_j \stackrel{\triangle}{=} f_0 + \sqrt{\frac{M \log M}{c_0 n}} g_j, \forall 1 \leqslant j \leqslant M \right).$$

Since we have that $\int_{\mathbb{R}^d} g_j(x)dx = 0$ and $M \log M \leqslant c_0 n$ the elements of $\mathcal{F}$ are probability densities on $\mathbb{R}^d$. Then the elements of $\mathcal{F}$ are uniformly bounded by L, thus we have that $\mathcal{F} \subset \mathcal{P}$. If we take $f \in \mathcal{F}$, then we have $\min_{1 \leqslant j \leqslant M} \|f_j - f\|_\pi = 0$. Thus it is sufficient to bound from below $\sup_{f \in \mathcal{F}} E_f^{\otimes n} (w(r_n^{-1} \|T_n - f\|_\pi^\alpha))$. For any $j, j' \in \{1, \ldots, M\}$ distinct, we have

$$||f_j - f_{j'}||_\pi = \left(\frac{L}{3}\right)^{1-\frac{1}{\pi}} M^{-\frac{1}{\pi}} \sqrt{\frac{M \log M}{c_0 n}}, \text{ if } \pi < \infty$$

and

$$||f_j - f_{j'}||_\infty = \frac{L}{3} \sqrt{\frac{M \log M}{c_0 n}}.$$

Thus

$$\begin{aligned} Q_1(\pi) &= \left(\frac{LM}{3}\right)^{1/2-1/\pi} \\ Q_2(\pi) &= \sqrt{\frac{LM}{3}} Q_1(\pi). \end{aligned}$$

Since $M \log M \leqslant c_0 n$, we have

$$||f_j - f_{j'}||_\pi = \sqrt{\frac{L}{3}} Q_1(\pi) \sqrt{\frac{\log M}{c_0 n}} \geqslant Q_2(\pi) \frac{\log M}{c_0 n}.$$

We have $||f_j - f_{j'}||_\pi \geqslant c r_n$ for any $j, j' \in \{1, \ldots, M\}$, $j \neq j'$, where c is an absolute constant.

Define $b_n = \sqrt{(M \log M)/(c_0 n)}$. For any $j, j' \in \{1, \ldots, M\}$, $j \neq j'$, we have

$$\begin{aligned} K(P_{f_j}^{\otimes n}, P_{f_{j'}}^{\otimes n}) &= nK(P_{f_j}, P_{f_{j'}}) \\ &= n \int_{[0,1]^d} \log\left(\frac{f_j(x)}{f_{j'}(x)}\right) f_j(x)dx \\ &= \frac{n}{2M}\left(\left(1 + \frac{b_n}{2}\right)\log\left(1 + \frac{b_n}{2}\right) + \left(1 - \frac{b_n}{2}\right)\log\left(1 - \frac{b_n}{2}\right)\right) \\ &+ \frac{n}{2M}\left(-\log\left(1 - \frac{b_n^2}{4}\right)\right) \\ &\leqslant \frac{n b_n^2}{4M}\left(1 + 2\sum_{k=1}^{\infty} \frac{1}{k 4^k}\right) \\ &\leqslant \left(\frac{1}{4} + \frac{\log(4/3)}{2})\right) \frac{1}{c_0} \log M \\ &\leqslant \frac{\log M}{16}, \text{ since } c_0 \geqslant 4(1 + 2\log(4/3)). \end{aligned}$$

We have used the inequality $\log x \leqslant x-1$, $\forall x > 0$, the equality $-\log(1-x) = \sum_{k=1}^{\infty} x^k/k$, $\forall x \in [-1, 1)$ and the fact that $b_n \leqslant 1$ in the third line. We use Lemma 8.1 to conclude. □

We obtain the lower bounds corresponding to the upper bounds given by the oracle inequalities (8.14) and (8.17) if we take respectively $w(x) = x$ and $w(x) = \mathbb{1}_{[c,\infty)}(x)$ for a constant $c > 0$.

It is interesting to compare Theorem 8.3 with [92, 89] where lower bounds for the aggregation of densities were derived under the L^2 risk. For $\pi = 2$ we have $Q_1(2) = 1$ and $Q_2(2) = \sqrt{LM/3}$ and the resulting lower bound in Theorem 8.3 is of the order $(L(\log M)/n)^{\alpha/2}$ as in [92, 89] with $\alpha = 2$. Considering the Gaussian white noise model, Goldenshluger [46] analyzed that for $\pi > 2$ the quantity $Q_1(\pi)$ becomes large. Therefore, Aggregation of arbitrary estimators is impossible when $\pi > 2$. We consider here the density estimation framework and we obtain similar oracle inequalities with the presence of the additional quantity $Q_2(\pi)$. Whereas $Q_1(\pi) \leqslant 1$ when $1 \leqslant \pi \leqslant 2$, the quantity $Q_2(\pi)$ can be large and Theorem 8.3 states that the upper bounds obtained in Section 8.3 cannot be improved. Thus aggregation of arbitrary estimators under the L^π risk is impossible with the Goldenshluger procedure even when $1 \leqslant \pi \leqslant 2$. However, in Section 8.4 we propose to aggregate a particular family of wavelet thresholded estimators. For this particular case, we prove that the quantities $Q_j(\pi)$, $j = 1, 2$, can be controlled nicely, thus leading to improvement of some known results on rate adaptive estimation.

8.3 The Goldenshluger-Lepski procedure

This procedure was proposed in [47] to aggregate particular classes of linear estimators such as wavelet estimators. It does not work for arbitrary class of estimators.

8.3.1 The aggregation procedure

Denote by

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i}$$

the empirical measure associated to the observations $\mathbb{X}_n$. Consider the following linear estimators:

$$\hat{f}_j(\cdot) = \int K_j(t, \cdot) dP_n(t), \, \forall j \in \mathcal{J},$$

where $\mathcal{J}$ is a finite subset of $\mathbb{N}$ and the kernels $K_j : \mathbb{R}^2 \to \mathbb{R}$ are integrable functions.

The next assumption is crucial to derive oracle inequality for this procedure.

Assumption 8.1. *The kernels K_j satisfy the following condition*

$$\int K_j(t,y)K_v(y,x)dy = K_{j\wedge v}(t,x), \quad \forall j,v \in \mathcal{J}.$$

We need another assumption on the kernels K_j.

Assumption 8.2. *The kernels K_j satisfy*

$$\max_{y\in\mathbb{R}} \int_{\mathbb{R}} |K_j(y,t)|dt \leqslant \overline{K} < \infty, \forall j \in \mathcal{J}.$$

Define the event

$$\mathcal{A}_\kappa = \bigcap_{j=1}^{n} \left\{ \|\hat{f}_j - E_f^{\otimes n}(\hat{f}_j)\|_\infty \leqslant \kappa\gamma(j) \right\}, \tag{8.18}$$

where $\gamma(j) = \gamma(j,n,\|f\|_\infty) > 0$ for any $j \in \mathcal{J}$. The quantities $(\gamma_j)_{j\in\mathcal{J}}$ are taken such that the event $\mathcal{A}_\kappa$ holds with high probability.

Define

$$\hat{B}_j \quad = \quad \frac{\max_{v\geqslant j, v\in\mathcal{J}} \left\{ \|\hat{f}_j - \hat{f}_v\|_\infty - 2\kappa\gamma(v) \right\}}{\overline{K}}, \forall j \in \mathcal{J}. \tag{8.19}$$

The model selection procedure is given by the following formulas:

$$\begin{aligned} \hat{\jmath} &= \arg\min_{j\in\mathcal{J}} \left\{ \hat{B}_j + 2\kappa\gamma(j) \right\} \\ \tilde{f}_n &= \hat{f}_{\hat{\jmath}}, \end{aligned} \tag{8.20}$$

where $\kappa > 0$ is a parameter.

8.3.2 Oracle inequality for the sup-norm

Define

$$B_j = E_f^{\otimes n}(\hat{f}_j) - \hat{f}_j, \forall j \in \mathcal{J}.$$

The following oracle inequality is a straightforward adaptation to the density estimation model of the result of [47] obtained in the Gaussian white noise model.

Theorem 8.4. *Assume that $\|f\|_\infty < \infty$ and that the estimators $(\hat{f}_j)_{j\in\mathcal{J}}$ are ordered such that the sequence $(\gamma(j))_{j\in\mathcal{J}}$ is non-decreasing. Let Assumptions 8.1-8.2 be satisfied. Then the estimator (8.19) satisfies on the event $\mathcal{A}_\kappa$*

$$\|\tilde{f}_n - f\|_\infty \leqslant (1 + 2(1\vee\overline{K})) \min_{j\in\mathcal{J}} \{\|B_j\|_\infty + 2\kappa\gamma(j,n,\|f\|_\infty)\}.$$

Proof. We have on the event $\mathcal{A}_\kappa$ for any integers $j, v \in \mathcal{J}$ such that $v \geqslant j$

$$\begin{aligned}\|\hat{f}_j - \hat{f}_v\|_\infty &\leqslant \|E_f^{\otimes n}(\hat{f}_j) - E_f^{\otimes n}(\hat{f}_v)\|_\infty + \|\hat{f}_j - E_f^{\otimes n}(\hat{f}_j)\|_\infty + \|\hat{f}_v - E_f^{\otimes n}(\hat{f}_v)\|_\infty \\ &\leqslant \|E_f^{\otimes n}(\hat{f}_j) - E_f^{\otimes n}(\hat{f}_v)\|_\infty + 2\kappa\gamma(v).\end{aligned}$$

Next, for any $v \geqslant j$

$$\begin{aligned}E_f^{\otimes n}(\hat{f}_j) - E_f^{\otimes n}(\hat{f}_v) &= \int (K_j(t,x) - K_v(t,x)) f(t) dt \\ &= \int K_v(y,x) \left(\int K_j(t,y) f(t) dt - f(y) \right) dy \\ &= \int K_v(y,x) B_j(y) dy,\end{aligned}$$

where we have used that $K_j(t,x) = K_{j \wedge v}(t,x) = \int K_j(t,y) K_v(y,x) dy$ for any $v \geqslant j$ and the Fubini Theorem in the second line. Assumption 8.2 yields

$$\begin{aligned}E_f^{\otimes n}(\hat{f}_j) - E_f^{\otimes n}(\hat{f}_v) &\leqslant \left(\int |K_v(y,x)| dy \right) \|B_j\|_\infty \\ &\leqslant \overline{K} \|B_j\|_\infty.\end{aligned}$$

Combining the above two displays yields that, on the event $\mathcal{A}_\kappa$,

$$\hat{B}_j \leqslant \|B_j\|_\infty, \, \forall j \in \mathcal{J}. \tag{8.21}$$

Next we have

$$\|\tilde{f}_n - f\|_\infty \leqslant \|\hat{f}_{\hat{\jmath}} - \hat{f}_{j^*}\|_\infty + \|\hat{f}_{j^*} - f\|_\infty.$$

For the second term on the RHS we have, on the event $\mathcal{A}_\kappa$,

$$\|\hat{f}_{j^*} - f\|_\infty \leqslant \|B_{j^*}\|_\infty + \kappa\gamma(j^*). \tag{8.22}$$

For the first term, we consider two cases. If $\hat{\jmath} \leqslant j^*$, then on the event $\mathcal{A}_\kappa$

$$\begin{aligned}\|\hat{f}_{\hat{\jmath}} - \hat{f}_{j^*}\|_\infty &= \|\hat{f}_{\hat{\jmath}} - \hat{f}_{j^*}\|_\infty - 2\kappa\gamma(j^*) + 2\kappa\gamma(j^*) \\ &\leqslant \overline{K} \hat{B}_{\hat{\jmath}} + 2\kappa\gamma(j^*).\end{aligned}$$

If $\hat{\jmath} > j^*$, then

$$\|\hat{f}_{\hat{\jmath}} - \hat{f}_{j^*}\|_\infty \leqslant \overline{K} \hat{B}_{j^*} + 2\kappa\gamma(\hat{\jmath}).$$

Thus we get

$$\begin{aligned}\|\hat{f}_{\hat{\jmath}} - \hat{f}_{j^*}\|_\infty &\leqslant \left(\overline{K} \hat{B}_{j^*} + 2\kappa\gamma(\hat{\jmath}) \right) \mathbb{1}_{\hat{\jmath} \leqslant j^*} + \left(\overline{K} \hat{B}_{\hat{\jmath}} + 2\kappa\gamma(j^*) \right) \mathbb{1}_{\hat{\jmath} > j^*} \\ &\leqslant (1 \vee \overline{K}) \left[\hat{B}_{j^*} + 2\kappa\gamma(j^*) + \hat{B}_{\hat{\jmath}} + 2\kappa\gamma(\hat{\jmath}) \right] \\ &\leqslant 2(1 \vee \overline{K}) \left[\hat{B}_{j^*} + 2\kappa\gamma(j^*) \right], \end{aligned} \tag{8.23}$$

where we have used the definition of $\hat{B}_j$ in the last line. Combining (8.21), (8.22) and (8.23) yields the result.

□

8.4 An application to rate adaptive density estimation

8.4.1 Wavelets, Besov spaces

We recall some well-known facts on wavelet expansions, see, e.g., the monograph [49]. Let $\phi \in L^2(\mathbb{R})$ be a father wavelet, i.e., the family $\{\phi(\cdot - k) : k \in \mathbb{Z}\}$ is an orthonormal system in $L^2(\mathbb{R})$, and the linear spaces $V_0 = \{f_x = \sum_k c_k \phi(x - k) : (c_k)_{k\in\mathbb{Z}} \in l_2\}$, $V_j = \{h_x = f(2^j x) : f \in V_0\}$, $j \geqslant 1$ are nested, that is $V_{j-1} \subset V_j$ for $j \in \mathbb{N}$, and their union is dense in $L^2(\mathbb{R})$: $\overline{\bigcup_{j=0}^{\infty} V_j} = L^2(\mathbb{R})$. We assume from now on that ϕ is bounded and compactly supported. Then the series

$$K(y, x) = \sum_{k\in\mathbb{Z}} \phi(y - k)\phi(x - k),$$

is a finite sum for any $x, y \in \mathbb{R}$ and

$$|K(y, x)| \leqslant \Phi(|y - x|), \quad \sup_{x\in\mathbb{R}} \sum_{k\in\mathcal{Z}} |\phi(x - k)| < \infty, \tag{8.24}$$

where $\Phi : \mathbb{R}^+ \to \mathbb{R}^+$ is bounded and compactly supported. Define

$$K_j(y, x) = 2^j K(2^j y, 2^j x),\ j \in \mathbb{N}.$$

For any $f \in L^p(\mathbb{R})$, $1 \leqslant p \leqslant \infty$, the projection of f onto V_j is

$$K_j(f)(y) = \int K_j(x, y) f(x) dx = \sum_{k\in\mathcal{Z}} 2^j \phi(2^j y - k) \int \phi(2^j x - k) f(x) dx, \quad \forall y \in \mathbb{R}.$$

Note that the above series converges pointwisely. If $f \in L^1(\mathbb{R})$, then the convergence of the series takes place in $L^p(\mathbb{R})$, $1 \leqslant p \leqslant \infty$. Let ψ be a mother wavelet associated to the father wavelet ϕ. Fix $j_0 \in \mathbb{N}$. The family $\{2^{j_0/2}\phi(2^{j_0}(\cdot) - k), 2^{l/2}\psi(2^l(\cdot) - k) : k \in \mathbb{Z}, l \in \mathbb{N}\}$ is an orthonormal basis of $L^2(\mathbb{R})$. Any function $f \in L^p(R)$ admits the wavelet decomposition

$$f(y) = \sum_k \alpha_{j_0,k}(f)\phi_{j_0,k}(y) + \sum_{j=j_0}^{\infty} \sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}(x),$$

where $\phi_{j_0,k}(y) = 2^{j_0/2}\phi(2^{j_0}y - k)$, $\psi_{j,k}(y) = 2^{j/2}\phi(2^j y - k)$, $\alpha_{j_0,k} = \int_{\mathbb{R}} f(x)\phi_{j_0,k}(x)dx$ and $\beta_{j,k} = \int f(x)\psi_{j,k}(x)dx$. The projection of f onto V_j is then given by

$$K_j(f)(y) = \sum_k \alpha_{j_0,k}\phi_{j_0,k}(y) + \sum_{l=j_0}^{\infty}\sum_{k=0}^{2^j-1} \beta_{l,k}\psi_{l,k}(y). \tag{8.25}$$

If ϕ and Ψ are bounded and compactly supported, then (8.25) holds pointwise. If $f \in L^1(\mathbb{R})$, then (8.25) also holds in $L^p(\mathbb{R}), 1 \leqslant p \leqslant \infty$.

The following properties are immediate consequences of the definition of the projections K_j when ϕ is bounded and compactly supported:

$$\int K_j(t,y)K_v(y,x)dy = K_{j\wedge v}(x),$$

since the spaces V_j are nested, and

$$\sup_j \sup_x \int |K_j(x,y)|dy \leqslant \sup_j \sup_x \int 2^j |\Phi(2^j|x-y|)|dy = \int_{\mathbb{R}^+} \Phi(|x|)dx.$$

Thus Assumptions 8.1 and 8.2 are satisfied with the constant $\overline{K} = \int_{\mathbb{R}^+} \Phi(|x|)dx$.

Assumption 8.3. *The father and mother wavelets ϕ and ψ have N first vanishing moments and N continuous derivatives for some integer $N \geqslant 1$.*

Let $L \in (0,\infty)$, $s \in (0,N)$, $p \in [1,\infty)$ et $r \in [1,\infty)$. In this setting, the Besov spaces $B^s_{p,r}$ are characterized by the behavior of the wavelet coefficients. In particular, we say that a function $f = \sum \alpha_{j_0,k}\phi_{j_0,k} + \sum \beta_{j,k}\psi_{j,k}$ belongs to the Besov ball $B^s_{p,r}(L)$ if and only if the associated wavelet coefficients satisfy

$$\left(\sum_{k=0}^{2^{j_0}-1} |\alpha_{j_0,k}|^p\right)^{1/p} + \left(\sum_{j=j_0-1}^{\infty}\left(2^{j(s+1/2-1/p)}\left(\sum_{k=0}^{2^j-1}|\beta_{j,k}|^p\right)^{1/p}\right)^r\right)^{1/r} \leqslant L.$$

We say that a function f belongs to the Besov ball $B^s_{p,\infty}(L)$ if and only if the associated wavelet coefficients satisfy

$$\left(\sum_{k=0}^{2^{j_0}-1} |\alpha_{j_0,k}|^p\right)^{1/p} + \sup_{j\geqslant j_0-1} 2^{j(s+1/2-1/p)}\left(\sum_{k=0}^{2^j-1}|\beta_{j,k}|^p\right)^{1/p} \leqslant L.$$

We say that a function f belongs to the Besov ball $B^s_{\infty,\infty}(L)$ if and only if the associated wavelet coefficients satisfy

$$\max_{0\leqslant k\leqslant 2^{j_0}-1} |\alpha_{j_0,k}| + \sup_{j\geqslant j_0-1} 2^{j(s+1/2)} \max_{0\leqslant k\leqslant 2^j-1} |\beta_{j,k}| \leqslant L.$$

Note that the above definitions are based on the wavelet coefficient characterization of the Besov spaces which holds true only under some finite moment conditions on the father wavelet ϕ and Φ (see Theorem 9.6 page 118 in [49]). These conditions are typically satisfied if ϕ is bounded, compactly supported and satisfies Assumption 8.3.

From now on, we consider the classes of functions:

$$\tilde{B}^s_{p,r}(L, L_0) = \{f \in B^s_{p,r}(L),\ \mathrm{supp}(f) \subset [-A, A],\ ||f||_\infty < L_0\},$$

where $\mathrm{supp}(f)$ denotes the support of f and A and $L_0 < \infty$ are given positive constants.

We assume that ψ is compactly supported on $\mathbb{R}$. It is known that compactly supported wavelet bases are unconditional bases for L^π spaces on $\mathbb{R}$ with $1 < \pi < \infty$ and satisfy the Temlyakov property (cf. [58] for more details and references). Define $I = I_{j_0} = \{(j,k) : j \geqslant j_0,\ k \in \{0, \ldots, 2^j - 1\}\}$.

1. *Property of unconditionality.* For any $1 < \pi < \infty$, there exist some constants c_π and C_π such that, for any subset $F \subset I$ and any sequence $u = (u_{j,k})_{(j,k)\in F}$ we have

$$c_\pi \left\| \sum_{(j,k)\in F} u_{j,k}\psi_{j,k} \right\|_\pi \leqslant \left\| \left(\sum_{(j,k)\in F} |u_{j,k}\psi_{j,k}|^2 \right)^{1/2} \right\|_\pi \leqslant C_\pi \left\| \sum_{(j,k)\in F} u_{j,k}\psi_{j,k} \right\|_\pi$$

2. *Temlyakov's property.* For any $0 < \pi < \infty$, there exist some constants c'_π and C'_π such that for any subset $F \subset I$ we have

$$c'_\pi \sum_{(j,k)\in F} ||\psi_{j,k}||^\pi_\pi \leqslant \left\| \left(\sum_{(j,k)\in F} |\psi_{j,k}|^2 \right)^{1/2} \right\|^\pi_\pi \leqslant C'_\pi \sum_{(j,k)\in F} ||\psi_{j,k}||^\pi_\pi.$$

These two properties will be crucial in the control of the quantity $Q_2(\pi)$ in the proof of Theorem 8.7.

8.4.2 Minimax wavelet estimators

The Minimax criterion

The L^π risk of any arbitrary estimator $\hat{f}_n$ based on the sample $\mathbb{X}_n$ over a functional space B is

$$R_{n,\pi}(\hat{f}_n, B) = \sup_{f\in B} E^{\otimes n}_f(||\hat{f}_n - f||^\pi_\pi),\ 1 \leqslant \pi < \infty.$$

The minimax L^π risk over B is defined by

$$r_n(B, L^\pi) = \inf_{\hat{f}_n} R_{n,\pi}(\hat{f}_n, B), \tag{8.26}$$

where the infimum is taken over all estimators $\hat{f}_n$ (measurable functions taking their values in a space containing B) of f. We recall that for two nonnegative sequences $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$, $a_n \asymp b_n$ means that

$$0 < \liminf_{n\to\infty} \frac{a_n}{b_n} \leqslant \limsup_{n\to\infty} \frac{a_n}{b_n} < \infty.$$

We state below Theorem 10.3 page 146 of [49]. This theorem gives the minimax risk on the classes $\tilde{B}^s_{p,r}(L, L_0)$.

Theorem 8.5. *Let $1 \leqslant p \leqslant \infty$, $1 \leqslant r \leqslant \infty$, $s > 1/p$ and $1 \leqslant \pi < \infty$. Let ϕ satisfy Assumption 8.3. Assume also that ϕ is bounded and compactly supported. Then*

$$r_n(\tilde{B}^s_{p,r}(L, L_0), L_\pi) \asymp \begin{cases} \left(\frac{1}{n}\right)^{\frac{\pi s}{1+2s}}, & \text{if } p > \frac{\pi}{1+2s}, \\ \left(\frac{\log n}{n}\right)^{\pi s'}, & \text{if } p < \frac{\pi}{1+2s}, \end{cases} \tag{8.27}$$

where $s' = \frac{\left(s-\frac{1}{p}+\frac{1}{\pi}\right)}{2(s-\frac{1}{p})+1}$. For the boundary case $p = \frac{\pi}{1+2s}$ there exist constants $0 < c < C$ and $\delta > 0$ such that for n large enough

$$c\left(\frac{\log n}{n}\right)^{\pi s'} \leqslant r_n(\tilde{B}^s_{p,r}(L, L_0), L_\pi) \leqslant C\left(\frac{\log n}{n}\right)^{\pi s'} (\log n)^\delta. \tag{8.28}$$

This theorem was proved in Donoho, Johnstone, Kerkyacharian and Picard [33]. They also proposed a wavelet thresholding procedure which is rate adaptive for the case $p < \pi/(1+2s)$ and adaptive up to a logarithmic factor if $p \geqslant \pi/(1+2s)$. Delyon and Juditsky [26] proposed a nonadaptive wavelet thresholding procedure achieving the minimax rates for the cases $p > \pi/(1+2s)$ and $p < \pi/(1+2s)$. For further details about the minimax L^π risks, see the monograph [49] where one can find more references.

Linear wavelet estimators

Consider the following estimator of the projection of f onto V_j

$$\begin{aligned} \hat{p}_n(y) = \hat{p}_n(y, j) &= \int K_j(y, x) dP_n \\ &= \sum_k \hat{\alpha}_{j_0,k}\phi_{j_0,k}(y) + \sum_{l=j_0}^{j}\sum_{k=0}^{2^l-1} \hat{\beta}_{l,k}\psi_{l,k}(x), \end{aligned}$$

where $\hat{\alpha}_{j_0,k} = \int \phi_{j_0,k}(x)dP_n(x)$, $\hat{\beta}_{l,k} = \int \psi_{l,k}(x)dP_n(x)$ are the empirical wavelet coefficients.

For the sake of completeness, we state below Theorem 3 of Giné and Nickl [43].

Theorem 8.6. *Fix an integer $N \geqslant 1$. Let $0 < s < N$ and Assumption 8.3 be satisfied. Assume also that ϕ and ψ are compactly supported. Consider the estimator $\hat{p}_n$ defined above where the the resolution $j = j^*$ is such that*

$$\left(\frac{n}{\log n}\right)^{\frac{1}{1+2s}} \leqslant 2^{j_*} \leqslant 2\left(\frac{n}{\log n}\right)^{\frac{1}{1+2s}},$$

Then, we have for n large enough

$$\sup_{f \in \tilde{B}^s_{\infty,\infty}(L,L_0)} E_f^{\otimes n}(||\hat{p}_n - f||_\infty) \leqslant C\left(\frac{\log n}{n}\right)^{\frac{s}{2s+1}},$$

where $C > 0$ is a constant depending only on L, L_0 and $\|\Phi\|_2$.

Proof. For the sake of simplicity, we denote by $\mathbb{E}$ the expectation $E_f^{\otimes n}$. We have

$$\begin{aligned} \|\hat{p}_n - f\|_\infty &\leqslant \|\mathbb{E}(\hat{p}_n) - f\|_\infty + \|\hat{p_n} - \mathbb{E}(\hat{p}_n)\|_\infty \\ &\leqslant \|K_j(f) - f\|_\infty + \|\hat{p_n} - \mathbb{E}(\hat{p}_n)\|_\infty, \end{aligned}$$

since $\mathbb{E}(\hat{p}_n) = K_j(f)$. Recall that ϕ and ψ satisfies Assumption 8.3. Since $f \in B^s_{\infty,\infty}(L)$ with $s < N$, we have

$$\|K_j(f) - f\|_\infty \leqslant C_1 2^{-js}, \tag{8.29}$$

where C_1 depends only on L. see, e.g., Theorem 9.4 page 117 in [49].

We treat now the variance term $\|\hat{p}_n - \mathbb{E}(\hat{p}_n)\|_\infty$. Define the random variables

$$Z_{j^*} = \sup_{y \in \mathbb{R}} |\hat{p}_n(y) - \mathbb{E}(\hat{p}_n(y))|,$$

and

$$Z_{j^*,i}(y) = \epsilon_i(K_{j^*}(X_i, y) - \mathbb{E}(K_{j^*}(X_i, y))), \quad \forall 1 \leqslant i \leqslant n.$$

For any i, y we have $\mathbb{E}(Z_{j^*,i}(y)) = 0$,

$$(Z_{j^*,i}(y)^2) \leqslant 2^{j^*}\|\Phi\|_2^2\|f\|_\infty \stackrel{\triangle}{=} \sigma^2_{j^*},$$

and

$$\sup_{y \in \mathbb{R}} |Z_{j^*,i}(y)| \leqslant 2^{j^*}\|\Phi\|_\infty \stackrel{\triangle}{=} U_{j^*}.$$

Define the set of functions

$$\mathcal{F}_{j^*} = \{K_{j^*}(\cdot, y) - \mathbb{E}(K_{j^*}(X_1, y)),\, y \in \mathbb{R}\}.$$

It is a known fact that the entropy number of the class $\mathcal{F}_{j^*}$ satisfies for any $\delta < U_{j^*}$

$$N(\mathcal{F}_{j^*}, L^2(Q), \delta) \leqslant \left(\frac{AU_{j^*}}{\delta}\right)^v,$$

where A, v depend only on Φ. See, e.g., Lemma 2 in [43] for compactly supported wavelets.

Inequality (38) in [45] yields

$$\mathbb{E}(Z_{j^*}) \leqslant 30\sqrt{2v}\sqrt{\frac{\sigma_{j^*}^2}{n}\log\frac{5AU_{j^*}}{\sigma_{j^*}}} + 15^2 2^5 v\frac{U_{j^*}}{n}\log\frac{5AU_{j^*}}{\sigma_{j^*}}.$$

The above inequality is obtained by standard arguments of the theory of empirical processes: Symmetrization, Dudley integral and Contraction principle.

The above display yields for n large enough that

$$\mathbb{E}\|\hat{p_n} - \mathbb{E}(\hat{p}_n)\|_\infty \leqslant C_2\sqrt{\frac{2^{j^*}j^*}{n}}, \tag{8.30}$$

where the constant $C_2 > 0$ depends only on L_0 and $\|\Phi\|_2$. Using (8.29) and (8.30) we get the result. □

In view of Theorems 8.5 and 8.6, the estimator $\hat{p}_n(\cdot, j^*)$ achieves the minimax rate of convergence. However, this estimator is not rate-adaptive since the choice of the optimum resolution j^* depends on the unknown regularity s. In Subsection 8.4.3 below, we exploit the aggregation procedure studied in Section 8.3 to build a minimax rate-adaptive estimator for the sup-norm.

Thresholded wavelet estimator

Consider the hard-thresholded estimator

$$\hat{p}_n^H(x) = \sum_k \hat{\alpha}_{j_0,k}(f)\phi_{j_0,k}(y) + \sum_{l=j_0}^{j_{\max}}\sum_{k=0}^{2^l-1}\hat{\beta}_{l,k}\mathbb{I}_{|\hat{\beta}_{lk}|>\lambda^{(l)}}\psi_{l,k}(x), \tag{8.31}$$

j_0 and $j_{\max}$ are integers satisfying

$$n^{1/(1+2N)} \leqslant 2^{j_0} \leqslant 2n^{1/(1+2N)},$$

and

$$n/\log n \leqslant 2^{j_{\max}} \leqslant 2n/\log n,$$

the thresholds $\lambda^{(j)} = \lambda_u^{(j)}$ $j = j_0, \ldots, j_{\max}$ are defined as follows

$$\lambda_u^{(j)} = \left(\frac{\rho(j-u)_+}{n}\right)^{1/2}, \quad j = j_0, \ldots j_{\max}, \tag{8.32}$$

where u satisfies

$$n^{1/(1+2s)} \leqslant 2^u \leqslant 2n^{1/(1+2s)}, \tag{8.33}$$

with s the regularity of the unknown density f and the parameter ρ is taken such that

$$\rho^2 \geqslant (4\log 2)(8\|f\|_\infty + (8\rho/(3\sqrt{2}))(||\psi||_\infty + \|f\|_\infty)). \tag{8.34}$$

We assume that n is large enough so that there exist integers j_0 and j_1 satisfying the above conditions.

This estimator was proposed by Delyon and Juditsky in [26]. They proved that for the choice (8.33) of the value of u depending on the unknown regularity s of the density f, the estimator achieves the minimax L_π rate without extra logarithmic factor for all configurations of (p, π, s) except the boundary case $p = \pi/(1 + 2s)$ where the estimator is suboptimal by a logarithmic factor. However, the estimator of [26] is nonadaptive since it requires the knowledge of the regularity s of the unknown density.

In Subsection 8.4.3 below, we propose to build a rate adaptive minimax estimator based on the above estimator and our aggregation procedure defined in Sections 8.2 and 8.3.

8.4.3 Rate adaptive minimax wavelet estimators

Rate adaptive estimation for the L^π-norm with $1 \leqslant \pi \leqslant 2$

We now define the estimator $\hat{f}_n$. We propose to use our aggregation procedure defined in Section 8.2 on a particular class of wavelet estimators defined below to build an adaptive estimator for the L^π risk with $1 \leqslant \pi \leqslant 2$. Consider two integers m and l such that $l = \left\lceil \frac{n}{\log n} \right\rceil$ and $m = n - l.$, and split the initial sample $\mathbb{X}_n$ into two independent subsamples $\mathbb{X}_m$ (training sample) and $\mathbb{X}_l$ (validation sample) respectively of size m and l. Set $M = \lceil \log m \rceil + 1$. With the training sample $\mathbb{X}_m$, we build the estimators f_u, $u = 1, \ldots, M$, as follows

$$f_u(x) = \sum_{k=0}^{2^{j_0}-1} \hat{\alpha}_{j_0,k}\phi_{j_0,k}(x) + \sum_{j=j_0}^{j_1}\sum_{k=0}^{2^j-1} \hat{\beta}_{j,k}\mathbb{I}_{|\hat{\beta}_{j,k}|\geqslant\lambda_u^{(j)}}\psi_{j,k}(x), \tag{8.35}$$

where

$$\hat{\alpha}_{j_0,k} = \frac{1}{m}\sum_{i=1}^{m}\phi_{j_0,k}(X_i), \quad \hat{\beta}_{j,k} = \frac{1}{m}\sum_{i=1}^{m}\psi_{j,k}(X_i), \tag{8.36}$$

j_0 and j_1 are integers satisfying

$$m^{1/(1+2N)} \leqslant 2^{j_0} \leqslant 2m^{1/(1+2N)},$$

and

$$m/(\log m) \leqslant 2^{j_1} \leqslant 2(m/(\log m)),$$

the thresholds $\lambda_u^{(j)}$ are

$$\lambda_u^j = \left(\frac{\rho(j-u+1)_+ \vee 1}{m} \right)^{1/2}, \ \ j = j_0, \dots j_1, \ u = 1, \dots, M, \tag{8.37}$$

where ρ is a parameter taken such that

$$\rho^2 \geqslant (4 \log 2)(8\|f\|_\infty + (8\rho/(3\sqrt{2}))(||\psi||_\infty + \|f\|_\infty)). \tag{8.38}$$

We assume that n is large enough so that there exist integers j_0 and j_1 satisfying the above conditions.

Now we freeze the subsample $\mathbb{X}_m$. Then $f_1, \dots, f_M$ are considered as deterministic functions. We fix $\pi \in [1,2]$ and we consider the set of functions Ψ given in (8.5) associated with $\mathcal{F} = \{f_1, \dots, f_M\}$. With the validation sample $\mathbb{X}_l$, we build the aggregate $\tilde{f}_l$ via the aggregation procedure (8.2)-(8.4). So our final estimator is

$$\hat{f}_n = \tilde{f}_l. \tag{8.39}$$

The wavelet estimator defined in (8.35)-(8.38) for a fixed integer u is a slight modification of the estimator of [26] where the authors considered (8.35) with $j_1 \asymp n/\log n$ and $\lambda_u^{(j)} = (\rho[(j-u)_+]/m)^{1/2}$. Inspection of the proof in [26] shows that our slightly modified version of the estimator of [26] still achieves the minimax L^π rate. But as we will see it in the proof, these modifications are necessary to control the terms $Q_j(\pi)$, $j = 1, 2$.

The estimator $\hat{f}_n$ is rate adaptive for the L^π risk with $\pi \in [1,2]$.

Theorem 8.7. *Fix $N \geqslant 1$ and $\pi \in [1,2]$. Let Assumption 8.3 be satisfied. We assume also that ϕ and ψ are compactly supported and that ψ satisfies the unconditional basis and Temlyakov properties. Consider the estimator $\hat{f}_n$ defined in (8.35)-(8.39). Then for all $s \in (1/p, N]$, $p \in [1, \infty]$, $r \in [1, \infty]$, $L, L_0 < \infty$ and n sufficiently large we have*

$$\sup_{f \in \tilde{B}^s_{p,r}(L,L_0)} E_f^{\otimes n}(||\hat{f}_n - f||_\pi^\pi) \leqslant \begin{cases} C\left(\frac{1}{n}\right)^{\frac{\pi s}{1+2s}}, & \text{if } p > \frac{\pi}{1+2s}, \\ C\left(\frac{\log n}{n}\right)^{\frac{\pi\left(s-\frac{1}{p}+\frac{1}{\pi}\right)}{2(s-\frac{1}{p})+1}}, & \text{if } p < \frac{\pi}{1+2s}, \\ C\left(\frac{\log n}{n}\right)^{\frac{\pi\left(s-\frac{1}{p}+\frac{1}{\pi}\right)}{2(s-\frac{1}{p})+1}} (\log n)^\delta, & \text{if } p = \frac{\pi}{1+2s}, \end{cases}$$

where δ is a positive constant depending only on π, s, p, r and $C > 0$ is a constant depending only on π, s, p, r, L, L_0.

As compared with Theorem 8.5, our procedure adaptively achieves the optimal rate on the scale of classes

$$\{\tilde{B}^s_{p,r}(L, L_0),\ (s,p,r) \in (1/p, N) \times [1,\infty] \backslash \{\pi/(1+2s)\} \times [1,\infty]\}.$$

Thus we improve upon the previously known results [23, 33, 59]. In particular, we cover the case $\pi/(1+2s) \leqslant p \leqslant \pi$ where the adaptive estimators of [33, 59] achieve logarithmically suboptimal rates. Chesneau and Lecué [23] constructed a rate adaptive estimator under the L^2 risk whereas our estimator $\hat{f}_n$ is rate adaptive under the L^π risk with $\pi \in [1,2]$. For the boundary case $p = \pi/(1+2s)$, our procedure is within a logarithmic factor of the minimax rate.

Proof. Assume w.o.l.g. that the functions $f_1, \ldots, f_M$ are distinct. Consider the set (8.5) of functions Ψ associated with $\mathcal{F} = \{f_1, \ldots, f_M\}$ where the estimators $f_u, u = 1, \ldots, M$ are defined in (8.35). Corollary 8.1 yields that

$$\begin{aligned} E_f^{\otimes l}\left(||\hat{f}_n - f||_\pi^\pi\right) \leqslant & C_1 \min_{1\leqslant u\leqslant M} ||f_u - f||_\pi^\pi \\ & + C_2\left(Q_2(\pi)\frac{\log M}{l}\right)^\pi + C_3\left(Q_1(\pi)\sqrt{\frac{L_0 \log M}{l}}\right)^\pi, \end{aligned} \tag{8.40}$$

where $E_f^{\otimes l}$ denotes the expectation w.r.t. the sample $\mathbb{X}_l$ and the constants $C_j, 1 \leqslant j \leqslant 3$, given in Corollary 1 depend only on π and

$$\begin{aligned} Q_1(\pi) &= \max_{1\leqslant u\neq u'\leqslant M} \frac{\|f_u - f_{u'}\|_{2\pi-2}^{\pi-1}}{\|f_u - f_{u'}\|_\pi^{\pi-1}} \\ Q_2(\pi) &= \max_{1\leqslant u\neq u'\leqslant M} \frac{\|f_u - f_{u'}\|_\infty^{\pi-1}}{\|f_u - f_{u'}\|_\pi^{\pi-1}}. \end{aligned}$$

For n large enough, we have that $(m/\log m)^{1/(1+2N)} \geqslant 1/2$, so that for any $s \in (1/p, N]$, there exists an integer u_s such that

$$\left(\frac{m}{\log m}\right)^{\frac{1}{1+2s}} \leqslant u_s \leqslant 2\left(\frac{m}{\log m}\right)^{\frac{1}{1+2s}}. \tag{8.41}$$

It follows from (8.41) that the integer u_s satisfies $u_s \leqslant M$ for any $s \in (1/p, N]$ since $m \geqslant 3$ for n large enough. As mentioned above, in an analogous way as in [26], it follows that if $f \in \tilde{B}^s_{p,r}(L, L_0)$ with $s \in (1/p, N]$ and if we choose ρ satisfying the relation (8.38) and the integer u_s satisfying (8.41), then the resulting estimator f_{u_s} defined in (8.35)-(8.38) is rate minimax on $\tilde{B}^s_{p,r}(L, L_0)$ if $p \neq \pi/(1+2s)$. So we have for m large enough that

$\min_{1\leqslant u\leqslant M} E_f^{\otimes m}[||f_u - f||_\pi^\pi] \leqslant cr_m(\tilde{B}^s_{p,r}(L,L_0),L_\pi)$ for any $f \in \tilde{B}^s_{p,r}(L,L_0)$, $s \in (1/p, N]$ where c is a constant depending only on p, s, π. Taking now the expectation w.r.t. the sample $\mathbb{X}_m$ in the oracle inequality (8.40), we get, for n large enough, that

$$\begin{aligned} E_f^{\otimes n}\left(||\hat{f}_n - f||_\pi^\pi\right) \leqslant & C_4 cr_m(\tilde{B}^s_{p,r}(L,L_0),L_\pi) \\ & + C_2 E_f^{\otimes m}\left(Q_1(\pi)^\pi\right)\left(L_0\sqrt{\frac{\log M}{l}}\right)^\pi \qquad (8.42) \\ & + C_3 E_f^{\otimes m}\left(Q_2(\pi)^\pi\right)\left(\frac{\log M}{l}\right)^\pi, \qquad (8.43) \end{aligned}$$

for any configuration (s,p,q,π) such that $p \neq \pi/(2s+1)$ where C_4 is a constant depending only on p, s, π, L, L_0.

Observe that $Q_1(\pi) \leqslant 1$ when $\pi \leqslant 2$, so that $E_f^{\otimes m}[Q_1(\pi)^\pi] \leqslant 1$. To control the value $E_f^{\otimes m}[Q_2(\pi)^\pi]$ in (8.42), we now show that the quantity $Q_2(\pi)\sqrt{(\log M)/l}$ is uniformly bounded by powers of $\log n$. If $\pi = 1$, then $Q_2(\pi) = 1$. If $\pi > 1$, we use the unconditional basis and Temlyakov's properties. Let $u < u'$ be integers in $\{1,\dots,M\}$. Define the set

$$\Delta_{u,u'} = \{(j,k),\, j_0 \leqslant j \leqslant j_1,\, 0 \leqslant k \leqslant 2^j - 1 \,:\, \lambda_{u'}^{(j)} \leqslant |\hat{\beta}_{j,k}| \leqslant \lambda_u^{(j)}\}.$$

We consider the case $1 < \pi \leqslant 2$. The unconditional property of the wavelet basis yields

$$||f_u - f_{u'}||_\pi \asymp \left|\left|\left(\sum_{(j,k)\in\Delta_{u,u'}} |\hat{\beta}_{j,k}\psi_{j,k}|^2\right)^{1/2}\right|\right|_\pi \triangleq I_{u,u'}.$$

Since the wavelet basis satisfies the Temlyakov property, we have

$$\sqrt{\frac{\rho}{m}}\left(\sum_{(j,k)\in\Delta_{u,u'}} ||\psi_{j,k}||_\pi^\pi\right)^{1/\pi} \leqslant I_{u,u'} \leqslant \sqrt{\frac{\rho j_1}{m}}\left(\sum_{(j,k)\in\Delta_{u,u'}} ||\psi_{j,k}||_\pi^\pi\right)^{1/\pi}.$$

We consider now the case $1 < \pi < \infty$. Define the sets

$$\mathcal{J}_{u,u'} = \{j \,:\, \exists k,\, (j,k) \in \Delta_{u,u'}\},$$

and for j fixed

$$\mathcal{K}_{j,u,u'} = \{k \,:\, (j,k) \in \Delta_{u,u'}\}.$$

Since ψ is compactly supported on $[-A', A']$ for some $A' > 0$, there exists a constant $C_7 > 0$ depending only on A' such that

$$||f_u - f_{u'}||_\infty \leqslant C_7 \sum_{j\in\mathcal{J}_{u,u'}} \max_{k\in\mathcal{K}_{j,u,u'}} |\hat{\beta}_{j,k}| 2^{j/2} ||\psi||_\infty. \qquad (8.44)$$

Then the unconditional basis and Temlyakov properties yield

$$\frac{||f_u - f_{u'}||_\infty}{||f_u - f_{u'}||_\pi} \leqslant C_8 \sqrt{j_1} \frac{\sum_{j \in \mathcal{J}_{u,u'}} 2^{j/2}}{\left(\sum_{(j,k) \in \Delta_{u,u'}} 2^{j(\pi/2-1)} \right)^{1/\pi}},$$

for some constant C_8 depending only on π, ψ. The Jensen inequality yields that

$$\frac{||f_u - f_{u'}||_\infty}{||f_u - f_{u'}||_\pi} \leqslant C_7 \sqrt{j_1} |\mathcal{J}_{u,u'}|^{(\pi-1)/\pi} \left[\frac{\sum_{j \in \mathcal{J}_{u,u'}} 2^{j\pi/2}}{\sum_{(j,k) \in \Delta_{u,u'}} 2^{j(\pi/2-1)}} \right]^{1/\pi}.$$

Since $|\mathcal{J}_{u,u'}| \leqslant j_1$ we have

$$\left[\frac{\sum_{j \in \mathcal{J}_{u,u'}} 2^{j\pi/2}}{\sum_{(j,k) \in \Delta_{u,u'}} 2^{j(\pi/2-1)}} \right]^{1/\pi} \leqslant c 2^{j_1/\pi},$$

where $c > 0$ is an absolute constant. This yields

$$Q_2(\pi) \leqslant C_9 2^{\frac{\pi-1}{\pi} j_1} j_1^{\frac{(\pi-1)^2}{\pi} + (\pi-1)/2},$$

for some constant C_9 depending only on π, ψ. Since $2^{j_1} \asymp m/(\log m)$ and $1 \leqslant \pi \leqslant 2$, we have, for n large enough, that

$$\begin{aligned} Q_2(\pi) \sqrt{\frac{\log M}{l}} &\leqslant C_9 2^{\frac{\pi-1}{\pi} j_1} j_1^{\frac{(\pi-1)^2}{\pi} + (\pi-1)/2} \sqrt{\frac{\log M}{l}} \\ &\leqslant C_9 \left(\frac{m}{\log m} \right)^{\frac{1}{2}} \left(\log \left(\frac{m}{\log m} \right) \right)^{(\pi-1)(\frac{\pi-1}{\pi} + \frac{1}{2})} \sqrt{\frac{\log((\log m) + 1)}{l}} \\ &\leqslant C_{10} (\log n)^\pi, \end{aligned}$$

where the constant C_{10} depends only on π, ψ. Combining the bounds on $Q_j(\pi)$, $j = 1, 2$ and (8.42) yields

$$\begin{aligned} E_f^{\otimes n} \left(||\tilde{f}_n - f||_\pi^\pi \right) \leqslant & C_4 r_m(\tilde{B}^s_{p,r}(L, L_0), L_\pi) \\ & + C_{11} \left(L_0 \frac{(\log n)^\pi \log(1 + (\log n))}{\sqrt{n}} \right)^\pi, \end{aligned} \tag{8.45}$$

where the constant C_{11} depends only on s, π, ψ. For n large enough, the second term on the right hand side of (8.45) is smaller than $r_m(\tilde{B}^s_{p,r}(L, L_0), L_\pi)$. We treat now the first term on the right hand side. For the case $p > \pi/(1 + 2s)$ we have that

$$\begin{aligned} r_m(\tilde{B}^s_{p,r}(L, L_0), L_\pi) &\leqslant C_{12} \left(\frac{1}{n(1 - 1/\log n)} \right)^{\frac{\pi s}{1+2s}} \\ &\leqslant C_{13} r_n(\tilde{B}^s_{p,r}(L, L_0), L_\pi) \left(1 - \frac{1}{\log n} \right)^{-\frac{\pi s}{1+2s}} \\ &\leqslant 2 C_{13} r_n(\tilde{B}^s_{p,r}(L, L_0), L_\pi), \end{aligned}$$

for n large enough, where the constants C_{12}, C_{13} depend only on p, s, π, ψ. Finally we use [26] to get the result. The cases $p < \pi/(1+2s)$ and $p = \pi/(1+2s)$ are treated similarly.

□

Rate adaptive estimation for the sup-norm

Let the integers $j_{\min}$ and $j_{\max}$ be such that

$$\left(\frac{n}{\log n}\right)^{1/(1+2N)} \leqslant 2^{j_{\min}} \leqslant 2\left(\frac{n}{\log n}\right)^{1/(1+2N)},$$

and

$$\frac{n}{\log n} \leqslant 2^{j_{\max}} \leqslant 2\frac{n}{\log n}.$$

Set $\mathcal{J} = [j_{\min}, j_{\max}] \cap \mathbb{N}$. Recall that the class of functions $\mathcal{F}_j = \{K_j(\cdot, y) - \mathbb{E}(K_j(X_1, y))\}$ is such that for any $\delta < U_j = 2^j \|\Phi\|_\infty$

$$N(\mathcal{F}_j, L^2(Q), \delta) \leqslant \left(\frac{AU_j}{\delta}\right)^v,$$

for any probability distribution Q, where A, v depend only on ϕ and $\sigma_j = 2^{j/2}\|\Phi\|_2 \|f\|_\infty$. Set

$$\begin{aligned} \gamma(j) = \gamma(j, n, \|f\|_\infty) &= 30\sqrt{2v}\sqrt{\frac{\sigma_j^2}{n}\log\frac{5AU_j}{\sigma_j}} + 15^2 2^5 v U_j \log\frac{5AU_j}{\sigma_j}, \\ \kappa &= 1 + N\log 2. \end{aligned}$$

Consider the class of linear estimators $(\hat{p}_n(\cdot, j))_{j\in\mathcal{J}}$. Recall that

$$\hat{B}_j = \frac{\max_{v \geqslant j,\, v\in\mathcal{J}} \left\{\|\hat{p}_n(\cdot, j) - \hat{p}_n(\cdot, v)\|_\infty - 2\kappa\gamma(v)\right\}}{\overline{K}},\ \forall j \in \mathcal{J}.$$

Define the aggregate $\tilde{f}_n$ as follows

$$\begin{aligned} \hat{\jmath} &= \arg\min_{j\in\mathcal{J}} \left\{\hat{B}_j + 2\kappa\gamma(j)\right\} \\ \tilde{f}_n &= \hat{p}_n(\cdot, \hat{\jmath}). \end{aligned} \tag{8.46}$$

We have the following result on $\tilde{f}_n$.

Theorem 8.8. *Fix $N \geqslant 1$. Let Assumption 8.3 be satisfied. We assume also that ϕ and ψ are compactly supported. Consider the estimator $\tilde{f}_n$ defined in (8.46). Then for all $0 < s < N$ and n large enough, we have*

$$\sup_{f\in\tilde{B}^s_{\infty,\infty}(L,L_0)} E_f^{\otimes n}(||\tilde{f}_n - f||_\infty) \leqslant C\left(\frac{\log n}{n}\right)^{\frac{s}{2s+1}},$$

where $C > 0$ is a constant depending only on L, L_0, $\|\Phi\|_2$ and N.

Giné and Nickl [43] proposed an estimator different from ours and proved its minimax rate adaptive property under the same assumptions as ours. The estimator of [43] is obtained by applying a model selection procedure close to Lepski's method. A significant difference from the original Lepski's method is that the biases are estimated by suprema of Rademacher processes in [44]. The construction of the estimator $\tilde{f}_n$ and the proof of our result are simpler since they do not involve such processes.

Proof. For the sake of simplicity, we denote respectively by $\mathbb{E}$ and $\mathbb{P}$ the expectation $E_f^{\otimes n}$ and the probability measure $P_f^{\otimes n}$. Recall that

$$\mathcal{A}_\kappa = \bigcap_{j\in\mathcal{J}} \left\{ \|\hat{p}_n(\cdot,j) - \mathbb{E}(\hat{p}_n(\cdot,j))\|_\infty \leqslant \kappa\gamma(j) \right\}.$$

For any $j \in \mathcal{J}$ set

$$B_j = \mathbb{E}(\hat{p}_n(\cdot,j)) - f, \qquad V_j = \mathbb{E}\|\hat{p}_n(\cdot,j) - \mathbb{E}(\hat{p}_n(\cdot,j))\|_\infty.$$

Theorem 8.4 yields on the event $\mathcal{A}_\kappa$

$$\|\tilde{f}_n - f\|_\infty \leqslant [1 + 2(1 \vee \overline{K})] \min_{j\in\mathcal{J}} \left\{ \|B_j\|_\infty + 2\kappa\gamma(j) \right\}. \tag{8.47}$$

On the event $\mathcal{A}_\kappa^c$ we have

$$\begin{aligned}
\mathbb{E}\|\tilde{f}_n - f\|_\infty &= \sum_{j\in\mathcal{J}} \mathbb{E}(\|\hat{p}_n(\cdot,j) - f\|_\infty \cdot \mathbb{1}_{\hat{j}=j}) \\
&\leqslant \sum_{j\in\mathcal{J}} \mathbb{E}(\|\hat{p}_n(\cdot,j) - f\|_\infty) \\
&\leqslant \sum_{j\in\mathcal{J}} V_j + \|B_j\|_\infty.
\end{aligned}$$

Applying Inequality (38) in [45] yields

$$\sum_{j\in\mathcal{J}} V_j \leqslant \sum_{j\in\mathcal{J}} c\sqrt{\frac{2^j j}{n}} + c'\frac{2^j j}{n},$$

for some constants $c, c' > 0$ depending only on $\|f\|_\infty$, $\|\Phi\|_2$ and $\|\Phi\|_\infty$. Simple computations yield

$$\sum_{j=j_{\min}}^{j_{\max}} V_j \leqslant c\sqrt{\frac{j_{\max}}{n}} \frac{2^{\frac{j_{\max}}{2}} - 2^{\frac{j_{\min}}{2}}}{\sqrt{2}-1} + c'\frac{j_{\max}}{n}(2^{j_{\max}} - 2^{j_{\min}}) \leqslant C, \tag{8.48}$$

for some constant $C > 0$ depending only on $\|f\|_\infty$, $\|\Phi\|_2$ and $\|\Phi\|_\infty$.

We have

$$\sum_{j=j_{\min}}^{j_{\max}} \|B_j\|_\infty \leqslant C' \sum_{j=j_{\min}}^{j_{\max}} 2^{-js} \leqslant \frac{C' 2^{-j_{\min} s}}{1-2^{-s}}, \tag{8.49}$$

where $C' > 0$ depends only on L and Φ. Next, we bound $\mathbb{P}(\mathcal{A}_\kappa^c)$ from above. Define

$$Z_j = \sup_{y \in \mathbb{R}} |\hat{p}_n(y,j) - \mathbb{E}(\hat{p}_n(y,j))|.$$

Recall that

$$\mathbb{E}(Z_j) \leqslant \gamma(j).$$

Applying Bousquet's version of Talagrand's concentration inequality, we get for any $x > 0$

$$\mathbb{P}\left(Z_j \geqslant \mathbb{E}(Z_j) + \sqrt{\frac{x}{n}(\sigma_j^2 + 2U_j \mathbb{E}(Z_j))} + \frac{U_j x}{3n} \right) \leqslant e^{-x}.$$

Simple computations yield for $x = 1 + N(\log 2)$

$$\mathbb{P}(Z_j \geqslant (1 + N(\log 2))\gamma) \leqslant 2^{-Nj}, \, \forall j \in \mathcal{J}.$$

We have by the union bound

$$\mathbb{P}(\mathcal{A}_\kappa^c) \leqslant \sum_{j \in \mathcal{J}} 2^{-Nj} \leqslant \frac{2^{-N j_{\min}}}{1-2^{-N}}. \tag{8.50}$$

Combining (8.48)-(8.50) yields

$$\mathbb{E}(\|\tilde{f}_n - f\|_\infty) \leqslant C \min_{j \in \mathcal{J}} \mathbb{E}(\|\hat{p}_n(\cdot, j) - f\|_\infty) + C' \left(\frac{\log n}{n} \right)^{\frac{N}{1+2N}},$$

where the constants C, C' are possibly different from the ones in (8.48)-(8.49) but depend only on L, L_0, Φ and N. Now It is sufficient to remark that for n large enough, there exists an integer $j^* \in \mathcal{J}$ such that $2^{j^*} \asymp n^{\frac{1}{1+2s}}$ and to combine the last display with Theorem 8.6. □

Bibliographie

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 2007. http ://www.springerlink.com/content/161105027v344n03.

[2] J.-Y. Audibert. A randomized online learning algorithm for better variance control. In *Learning theory*, volume 4005 of *Lecture Notes in Comput. Sci.*, pages 392–407. Springer, Berlin, 2006.

[3] F. Bach. Bolasso : Model consistent lasso estimation through the bootstrap. In *Proceedings of the 25 th International Conference on Machine Learning, Helsinki, Finland*, 2008.

[4] F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9 :1179–1225, 2008.

[5] P.L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48 :85–113, 2002.

[6] P.J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37(4) :1705–1732, 2009.

[7] L. Birgé. Model selection via testing : an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3) :273–325, 2006. ISSN 0246-0203.

[8] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[9] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3) : 203–268, 2001. ISSN 1435-9855.

[10] J.M. Borwein and A.S. Lewis. *Convex Analysis And Nonlinear Optimization : Theory And Examples*. Springer, 2006.

[11] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6) :495–500, 2002. ISSN 1631-073X.

[12] F. Bunea. Consistent selection via the Lasso for high dimensional approximating regression models. In *Pushing the limits of contemporary statistics : contributions in honor of Jayanta K. Ghosh*, volume 3 of *Inst. Math. Stat. Collect.*, pages 122–137. Inst. Math. Statist., Beachwood, OH, 2008.

[13] F. Bunea and A. Nobel. Sequential procedures for aggregating arbitrary estimators of a conditional mean. *IEEE Trans. Inform. Theory*, 54(4) :1725–1735, 2008. ISSN 0018-9448.

[14] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1 :169–194, 2007.

[15] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35 :1674–1697, 2007.

[16] E. Candes and T. Tao. The Dantzig selector : Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6) :2313–2351, 2005.

[17] O. Catoni. Universal aggregation rules with exact bias bounds. Technical report, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 et Paris 7, 1999.

[18] O. Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. ISBN 3-540-22572-2. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.

[19] G. Cavallanti, N. Cesa-Bianchi, and C. Gentile. Linear algorithms for online multitask classification. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, 2008.

[20] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, Cambridge, 2006. ISBN 978-0-521-84108-5 ; 0-521-84108-9.

[21] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1) :129–159 (electronic), 2001. ISSN 0036-1445. Reprinted from SIAM J. Sci. Comput. **2**0 (1998), no. 1, 33–61 (electronic) [MR1639094 (99h :94013)].

[22] C. Chesneau and M. Hebiri. Some theoretical results on the grouped variable lasso. http ://hal.archives-ouvertes.fr/hal-00145160/fr/. *Mathematical Journal of statistics*, 17(4) :317–326, 2007.

[23] C. Chesneau and G. Lecué. Adapting to unknown smoothness by aggregation of thresholded wavelet estimators. *Statist. Sinica*, 19(4), 2009.

[24] A.S. Dalalyan and A.B. Tsybakov. Pac-bayesian bounds for the expected error of aggregation by exponential weights. Technical report, Université Paris 6, CREST and CERTIS, Ecole des Ponts ParisTech, 2009. personal communication.

[25] A.S. Dalalyan and A.B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 97–111. Springer, Berlin, 2007.

[26] B. Delyon and A. Juditsky. On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.*, 3(3) :215–228, 1996. ISSN 1063-5203.

[27] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 1998. ISBN 0-387-98406-2.

[28] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York, 2001. ISBN 0-387-95117-2.

[29] P. Diggle. *Analysis of Longitudinal Data*. Oxford University Press, 2002.

[30] D.L. Donoho. For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, 59(6) :797–829, 2006. ISSN 0010-3640.

[31] D.L. Donoho and J. Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA*, 102(27) :9452–9457 (electronic), 2005. ISSN 1091-6490.

[32] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage : asymptopia ? *J. Roy. Statist. Soc. Ser. B*, 57(2) :301–369, 1995. ISSN 0035-9246. With discussion and a reply by the authors.

[33] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2) :508–539, 1996. ISSN 0090-5364.

[34] D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, 52(1) :6–18, 2006.

[35] C. Dossal. A necessary and sufficient condition for exact recovery by l_1 minimization. Technical report, Institut de mathématiques de Bordeaux, Université Bordeaux 1, 2008.

[36] L. Dümbgen, S.A. van de Geer, and J.A. Wellner. Nemirovski's Inequalities Revisited. Available on Arxiv, 2008.

[37] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2) :407–499, 2004. ISSN 0090-5364. With discussion, and a rejoinder by the authors.

[38] T. Evgeniou, M. Pontil, and O. Toubia. A convex optimization approach to modeling consumer heterogeneity in conjoint estimation. *Marketing Science*, 26 :805–818, 2007.

[39] W. Feller. *An introduction to probability theory and its applications. Vol. I.* Third edition. John Wiley & Sons Inc., New York, 1968.

[40] D. Foster and E. George. The risk inflation criterion for multiple regression. *Ann. Statist.*, 22 :1947–1975, 1994.

[41] I.E. Franck and J.H. Friedman. A statistical point of view of some chemometrics regression tools (with discussion). *Technometrics*, 35 :109–148, 1993.

[42] J. Fuchs. Recovery of exact sparse representations in the presence of bounded noise. Technical report, Irisa, Université de Rennes, 2004.

[43] E. Giné and R. Nickl. Uniform limit theorems for wavelet density estimators. *Annals of Probability*. to appear.

[44] E. Giné and R. Nickl. Adaptive estimation of the distribution function and its density in sup-norm loss by wavelet and spline projections. Technical report, University of Connecticut and University of Cambridge, 2009.

[45] E. Giné and R. Nickl. Uniform central limit theorems for kernel density estimators. *Probab. Theory Related Fields*, 141(3-4) :333–387, 2008. ISSN 0178-8051.

[46] A. Goldenshluger. A universal procedure for aggregating estimators. *Ann. Statist.*, 37 (1) :542–568, 2009.

[47] A. Goldenshluger and O. Lepski. Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probab. Theory Related Fields*, 143(1-2) :41–71, 2009. ISSN 0178-8051.

[48] E. Greenshtein and Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6) :971–988, 2004. ISSN 1350-7265.

[49] W. Härdle, G. Kerkyacharian, D. Picard, and A.B. Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1998. ISBN 0-387-98453-4.

[50] D. Haussler, J. Kivinen, and M.K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Trans. Inform. Theory*, 44(5) :1906–1925, 1998. ISSN 0018-9448.

[51] M. Hebiri. *Quelques Questions De Sélection De Variables Autour De l'Estimateur Lasso*. PhD thesis, Université Paris Diderot - Paris 7, 2009.

[52] N. Hengartner and M. Wegkamp. Estimation and selection procedures in regression : an L_1 approach. *Canad. J. Statist.*, 29(4) :621–632, 2001. ISSN 0319-5724.

[53] C. Hsiao. *Analysis of Panel Data*. Cambridge University Press, 2003.

[54] J. Huang, J.L. Horowitz, and F Wei. Variable selection in nonparametric additive models. manuscript. Technical report, Department of Economics, Northwestern University, 2008.

[55] G.M. James and P. Radchenko. A generalized dantzig selector with shrinkage tuning. *Biometrika*, 2008.

[56] A. Juditsky and A. Nemirovski. On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization. *submitted to Mathematical Programming Series B., Special Issue on Machine Learning*, 2008. arXiv :0809.2650v1.

[57] A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *Ann. Statist.*, 36(5) :2183–2206, 2008. ISSN 0090-5364.

[58] G. Kerkyacharian and D. Picard. Thresholding algorithms, maxisets and well-concentrated bases. *Test*, 9(2) :283–344, 2000. ISSN 1133-0686. With comments, and a rejoinder by the authors.

[59] G. Kerkyacharian, D. Picard, and K. Tribouley. L^p adaptive density estimation. *Bernoulli*, 2(3) :229–247, 1996. ISSN 1350-7265.

[60] J. Kivinen and M.K. Warmuth. Averaging expert predictions. In *Computational learning theory (Nordkirchen, 1999)*, volume 1572 of *Lecture Notes in Comput. Sci.*, pages 153–167. Springer, Berlin, 1999.

[61] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Ann. Statist.*, 28(5) : 1356–1378, 2000. ISSN 0090-5364.

[62] V. Koltchinskii. Sparsity in penalized risk minimization. *Ann. IHP*, to appear.

[63] V. Koltchinskii. Dantzig selector and sparsity oracle inequalities. *Bernoulli*, to appear.

[64] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In *Conference on Learning Theory, COLT-2008*, pages 229–238, 2008.

[65] G. Lecué. Lower bounds and aggregation in density estimation. *J. Mach. Learn. Res.*, 7 :971–981, 2006. ISSN 1532-4435.

[66] M. Ledoux and M. Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. ISBN 3-540-52013-9. Isoperimetry and processes.

[67] P. J. Lenk, W. S. DeSarbo, P. E. Green, and M. R. Young. Hierarchical Bayes conjoint analysis : recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, 15(2) :173–191, 1996.

[68] G. Leung and A.R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8) :3396–3410, 2006. ISSN 1557-9654.

[69] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient $l_{2,1}$-norm minimization. *COLT (electronic)*, 2009.

[70] K. Lounici. Generalized mirror averaging and D-convex aggregation. *Math. Methods Statist.*, 16(3) :246–259, 2007. ISSN 1066-5307.

[71] K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2 :90–102, 2008.

[72] K. Lounici. High-dimensional stochastic optimization with the generalized dantzig selector. Technical report, Laboratoire de Probabilités et Modèles aléatoires, Université Paris 7 and CREST, 2009. Submitted.

[73] K. Lounici. Aggregation of density estimators for the l^p-risk, $1 \leqslant p \leqslant \infty$. adaptive wavelet thresholded estimators. Technical report, Laboratoire de Probabilités et Modèles aléatoires, Université Paris 7 and CREST, 2009. work in progress.

[74] K. Lounici, M. Pontil, A.B. Tsybakov, and S.A. van de Geer. Taking advantage of sparsity in multi-task learning. *COLT (electronic)*, 2009.

[75] G. Lugosi and M. Wegkamp. Complexity regularization via localized random penalties. *Ann. Statist.*, 32(4) :1679–1697, 2004. ISSN 0090-5364.

[76] A. Maurer. Bounds for Linear Multi-Task Learning. *The Journal of Machine Learning Research*, 7 :117–139, 2006.

[77] D.A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (Santa Cruz, CA, 1999)*, pages 164–170 (electronic), New York, 1999. ACM.

[78] L. Meier, S.A. van de Geer, and P. Bühlmann. The Group Lasso for Logistic Regression. *Journal of the Royal Statistical Society, Series B*, 70(1) :53–57, 2006.

[79] L. Meier, S. van de Geer, and P. Buhlmann. High-dimensional additive modeling. arxiv :0806.4115. *Annals of Statistics*, to appear.

[80] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3) :1436–1462, 2006. ISSN 0090-5364.

[81] N. Meinshausen and Bühlmann P. Stability selection. Technical report, University of Oxford and ETH Zürich, 2008.

[82] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1) :246–270, 2009.

[83] Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2 :605–633, 2008.

[84] A. Nemirovski. Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, volume 1738 of *Lecture Notes in Math.*, pages 85–277. Springer, Berlin, 2000.

[85] A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization.* A Wiley-Interscience Publication. John Wiley & Sons Inc., New York, 1983. ISBN 0-471-10345-4. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[86] G. Obozinski, M.J. Wainwright, and M.I. Jordan. Union support recovery in high-dimensional multivariate regression. Technical report, Berkeley Stat., 2008.

[87] M.R. Osborne, B. Presnell, and B.A. Turlach. On the lasso and its dual. *J. Comput. Graph. Statist.*, 9(2) :319–337, 2000. ISSN 1061-8600.

[88] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. Spam : Sparse additive models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 22, 2007.

[89] Ph. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3) :260–280, 2007. ISSN 1066-5307.

[90] Ph. Rigollet and P. Zhao. Universal mirror averaging. Personal communication, 2006.

[91] M. Rosenbaum and A.B. Tsybakov. Sparse recovery under matrix uncertainty. *Ann. Statist., submitted*, 2009.

[92] A. Samarov and A.B. Tsybakov. Aggregation of density estimators and dimension reduction. In *Advances in statistical modeling and inference*, volume 3 of *Ser. Biostat.*, pages 233–251. World Sci. Publ., Hackensack, NJ, 2007.

[93] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2) :461–464, 1978. ISSN 0090-5364.

[94] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1) :267–288, 1996. ISSN 0035-9246.

[95] A.B. Tsybakov. A note on the lasso. Personal communication, 2008.

[96] A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1) :135–166, 2004. ISSN 0090-5364.

[97] A.B. Tsybakov. Optimal rates of aggregation. In *Computationnal Learning theory and Kernel Machines (COLT)*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 303–313. Springer, heidelberg, 2003.

[98] Alexandre B. Tsybakov. *Introduction à l'estimation non-paramétrique*, volume 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2004. ISBN 3-540-40592-5.

[99] S.A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2) :614, 2008.

[100] S.A. van de Geer. The deterministic lasso. Technical report, Seminar für Statistik ETH, Zürich, 2007.

[101] A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94640-3.

[102] V. Vovk. Aggregating strategies. In *3rd Annual Workshop on Computationnal Learning theory*, pages 372–383. Morgan Kaufman, San Mateo, 1990.

[103] M. Waingwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (lasso). *IEEE*, 2009.

[104] M.J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using l_1-constrained quadratic programming. Technical report, Department of Statistics, UC Berkeley, 2006.

[105] D.L. Wallace. Bounds for normal approximations of student's t and the chi-square distributions. *Ann. Math. Statist.*, 30 :1121–1130, 1959.

[106] M. Wegkamp. Model selection in nonparametric regression. *Ann. Statist.*, 31(1) : 252–273, 2003. ISSN 0090-5364.

[107] M. Wegkamp. Lasso type classifiers with a reject option. *Electron. J. Stat.*, 1 :155–168 (electronic), 2007. ISSN 1935-7524.

[108] J.M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data.* MIT Press, 2002.

[109] Y. Yang. Combining different procedures for adaptive regression. *J. Multivariate Anal.*, 74(1) :135–161, 2000. ISSN 0047-259X.

[110] Y. Yang. Aggregating procedures for a better performance. *Bernoulli*, 10 :25–47, 2004.

[111] Y.G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Ann. Statist.*, 13(2) :768–774, 1985. ISSN 0090-5364.

[112] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 68(1) :49–67, 2006.

[113] A. B. Yuditskiĭ, A. V. Nazin, A. B. Tsybakov, and N. Vayatis. Recursive aggregation of estimators by the mirror descent method with averaging. *Problemy Peredachi Informatsii*, 41(4) :78–96, 2005. ISSN 0555-2923.

[114] C.H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36(4) :1567–1594, 2008. ISSN 0090-5364.

[115] T. Zhang. From ϵ-entropy to KL-entropy : analysis of minimum information complexity density estimation. *Ann. Statist.*, 34(5) :2180–2210, 2006. ISSN 0090-5364.

[116] P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7 :2541–2563, 2006. ISSN 1532-4435.

[117] H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101 (476) :1418–1429, 2006. ISSN 0162-1459.

Résumé : Dans cette thèse nous traitons deux sujets. Le premier sujet concerne l'apprentissage statistique en grande dimension, i.e. les problèmes où le nombre de paramètres potentiels est beaucoup plus grand que le nombre de données à disposition. Dans ce contexte, l'hypothèse généralement adoptée est que le nombre de paramètres intervenant effectivement dans le modèle est petit par rapport au nombre total de paramètres potentiels et aussi par rapport au nombre de données. Cette hypothèse est appelée "*sparsity assumption*". Nous étudions les propriétés statistiques de deux types de procédures : les procédures basées sur la minimisation du risque empirique muni d'une pénalité l_1 sur l'ensemble des paramètres potentiels et les procédures à poids exponentiels. Le second sujet que nous abordons concerne l'étude de procédures d'agrégation dans un modèle de densité. Nous établissons des inégalités oracles pour la norme L^π, $1 \leqslant \pi \leqslant \infty$. Nous proposons ensuite une application à l'estimation minimax et adaptative en la régularité de la densité.

Mots-clés : Inégalités d'oracle, optimisation stochastique, agrégation, apprentissage statistique, grande dimension, sparsité, sélection de variables, Lasso, Dantzig Selector, estimation adaptative minimax.

Discipline : **Mathématiques**

Abstract : We treat two subjects. The first subject is about statistical learning in high-dimension, that is when the number of paramaters to estimate is larger than the sample size. In this context, the generally adopted assumption is that the number of true parameters is much smaller than the number of potential paramaters. This assumption is called the "*sparsity assumption*". We study the statistical properties of two types of procedures : the penalized risk minimization procedures with a l_1 penalty term on the set of potential parameters and the exponential weights procedures. The second subject is about the study of two aggregation procedures in a density estimation problem. We establish oracle inequalities for the L^π norm, $1 \leqslant \pi \leqslant \infty$. Next, we exploit these results to build minimax rate adaptive estimators of the density.

Key words : Oracle inequalities, stochastic optimization, aggregation, statistical learning, high dimension, sparsity, variable selection, Lasso, Dantzig Selector, minimax rate adaptive estimation.

Laboratoire de Probabilités et Modèles Aléatoires,
CNRS-UMR 7599, UFR de Mathématiques, case 7012
Université Paris 7, Denis Diderot
2, place Jussieu, 75251 Paris Cedex 05.